Design research in statistics education
On symbolizing and computer tools

Bakker, Arthur

Design research in statistics education: On symbolizing and computer tools / A. Bakker – Utrecht: CD-β Press, Center for Science and Mathematics Education – (CD-β wetenschappelijke bibliotheek; nr. 50; 2004)
Dissertation Utrecht University. – With references. – With a summary. – Met een samenvatting in het Nederlands.

ISBN 90-73346-58-4
Subject headings: mathematics education, design research, technology, semiotics, history of statistics

Cover: Zuidam & Uithof, Utrecht
Press: Wilco, Amersfoort

Design research in statistics education
On symbolizing and computer tools

Ontwikkelingsonderzoek in het statistiekonderwijs
Over symboliseren en computertools

(met een samenvatting in het Nederlands)

PROEFSCHRIFT
TER VERKRIJGING VAN DE GRAAD VAN DOCTOR
AAN DE UNIVERSITEIT UTRECHT,
OP GEZAG VAN DE RECTOR MAGNIFICUS PROF. DR. W. GISPEN
INGEVOLGE HET BESLUIT VAN HET COLLEGE VOOR PROMOTIES
IN HET OPENBAAR TE VERDEDIGEN
OP
MAANDAG 24 MEI 2004
DES MIDDAGS TE 14:30 UUR

door
Arthur Bakker
geboren op 3 januari 1970, te Hilversum

Promotoren:

Prof. dr. K.P.E. Gravemeijer	Faculteit Sociale Wetenschappen, Universiteit Utrecht Faculteit Wiskunde en Informatica, Universiteit Utrecht
Prof. dr. G. Kanselaar	Faculteit Sociale Wetenschappen, Universiteit Utrecht
Prof. dr. J. de Lange	Faculteit Wiskunde en Informatica, Universiteit Utrecht

The research reported in this dissertation was funded by the Netherlands Organization for Scientific Research (NWO) under grant number 575-36-003B.

Preface

I have always enjoyed reading prefaces of dissertations because they often reveal something about the person behind the text and about the litany of people involved in the research.

Although I was unable to formulate it in this way as a youngster, I have always been intrigued by how people think and learn. At school I read histories of philosophy to see where ideas came from, I learned to play the violin (my mother being a violinist and music therapist), psychology sounded interesting, and artificial intelligence was 'hot'. However, due to my father (a mathematics teacher) and the Mathematics and Pythagoras Olympiads, mathematics eventually caught my imagination as the purest form of human thinking. I chose to study mathematics, but its relations to other disciplines such as philosophy, logic, history, and education continually demanded attention. After receiving a Master's degree in mathematics, investigating the philosophical foundations of set theory proved rewarding, but this pursuit seemed somehow irrelevant to society. And while teaching secondary school mathematics felt very relevant, it was not my calling.

In 1998, I acquired a graduate position (*onderzoeker in opleiding*) at the Freudenthal Institute and mathematics education turned out to be the interdisciplinary mix I had been looking for without knowing it. This mix is reflected in my thesis: I studied the history of statistics to understand how statistical concepts evolved and to gain insights into promising ways to teach them. Conducting the design research required incorporating insights from educational science, psychology, mathematics and statistics education, and practical experience as a teacher. Last, the semiotic analyses reflect my interest in the philosophy of language.

First, I would like to express my gratitude to Koeno Gravemeijer, who succeeded in raising the funds for this project from the Netherlands Organization for Scientific Research (NWO). He coached me along the route of becoming a researcher in mathematics and statistics education. Gellof Kanselaar and Jan de Lange complemented the advisory team, each in their own helpful way. For more specialized topics I consulted experts who were forthcoming in their commentary on various chapters: Paul Cobb (Chapters 2, 4, and 8), Stephen Stigler (4), Cliff Konold (2, 4), and Michael Hoffmann (6, 8, 9). Thank you, Paul, Cliff, and Michael, for rewarding meetings and your hospitality as well. I also like to mention the inspiration I received from the Forum for Statistical Reasoning, Thinking, and Literacy, chaired by Dani Ben-Zvi and Joan Garfield.

Conducting Ph.D. research can be a solitary enterprise, but I was happy to participate in a larger project (*aandachtsgebied*). In this way, I was always able to discuss issues with and feel support from Monique Pijls, Dirk Hoek, Michiel Doorman, and in particular Paul Drijvers. The Freudenthal Institute formed another stimulating commu-

nity of practice, in which I learned a lot just by participating. In particular I would like to thank two colleagues, Mieke Abels and Corine van den Boer, who also taught during the teaching experiments, and their students. Aad Goddijn and Martin Kindt were always willing to discuss mathematical and historical issues.
The ways in which I have been assisted are numerous. Petra van Loon, Harm Bergevoet, Carolien de Zwart, Sofie Goemans, and Yan Wei Zhou assisted during the teaching experiments by interviewing and videotaping. Han Hermsen taught me how to use Framemaker, the word processor with which this book has been made. Tanja Klooster, Ellen Hanepen, and Parul Slegers helped to transcribe students' utterances. Anneleen Post, among others, helped me when Repetitive Strain Injury (or Carpal Tunnel Syndrome) hindered my work at the computer. Nathalie Kuijpers and Tim Muentzer corrected my English and Betty Heijman and Sylvia Eerhart assisted in the book's publication. And of course I would like to thank Frans van Galen with whom I have shared my office all these years. Huub van Baar and Yolande Jansen supported me as friends and ushers (*paranimfen*). Thank you all!
Jantien, you helped me in various ways and I am grateful for your very existence in my life. That is why I dedicate this book to you.

Table of Contents

1 Introduction

Information is the fuel of the knowledge society in which we live.
Johan van Benthem

The present study is a sequel to design research in statistics education carried out by Cobb, McClain, Gravemeijer, and their team in Nashville, TN, USA. The research presented in this thesis is also part of a larger research project on the role of IT in secondary mathematics education.[1] In the remainder of this introductory chapter we discuss the notions of the title *Design research in statistics education: on symbolizing and computer tools,* and identify the purpose of the research.

1.1 Statistics education

Statistics is seen as a science of variability and as a way to deal with the uncertainty that surrounds us in our daily life, in the workplace, and in science (Kendall, 1968; Moore, 1997). In particular, statistics is used to describe and predict phenomena that require collections of measurements. But what are the skills essential to the navigation of today's technological and information-laden society? Statistical literacy is one of those skills. Gal (2002) characterizes it as the ability to interpret, critically evaluate, and communicate about statistical information and messages. We give three instances to exemplify how citizens of modern society need at least some statistical literacy.

1 Many newspapers present graphs or data on the front page. Apparently, citizens are expected to understand and appreciate such condensed information; it is not just the educated who are confronted with statistical information. Research in statistics education, however, shows that graphs are difficult to interpret for most people:

> The increasingly widespread use of graphs in advertising and the news media for communication and persuasion seems to be based on an assumption, widely contradicted by research evidence in mathematics and science education, that graphs are transparent in communicating their meaning. (Ainley, 2000, p. 365)

It could also be that newspapers attempt to create a reliable or scientific impression.

2 More and more large companies have a policy of teaching almost all employees some basic statistics. This is often part of a quality control method; for instance,

1. Among the publications of the other project members are Drijvers (2003), Doorman (in press), Hoek, Seegers, & Gravemeijer (in press), and Pijls, Dekker, and Van Hout-Wolters (2003).

Six Sigma aims to increase profitability by controlling variation in production processes (e.g. Pyzdek, 2001). The basic idea of statistical process control is that variation and the chance of mistakes should be minimized and that to achieve that, every employee should be familiar with variation, usually measured by the standard deviation, around a target value, usually the mean (Descamps, Janssens, & Vanlangendonck, 2001). This makes statistics an instrument for economic success.

3 In almost every political and economic decision, at least some statistical information is used. Fishermen, for instance, negotiate with the government and perhaps environmental groups about fish quotas, which are based on data and statistical models (Van Densen, 2001). This makes statistics a language of power.

If we want to provide all students some basic statistical baggage, we need to teach statistical data analysis to school-aged children. Statistical literacy, however, is not an achievement that is readily established: the growing body of research in this area shows how much effort it takes to teach and learn statistical reasoning, thinking, and literacy (Garfield & Ahlgren, 1988; Shaughnessy, Garfield, & Greer, 1996). Students need to master difficult concepts and use complicated graphs, and teachers often lack the statistical background to help students do so (Makar & Confrey, in press; Mickelson & Heaton, in press). This implies that students need early exposure to statistical data analysis and that we need to know more about how to support them.

Besides societal need, there are also theoretical reasons to do research in statistics education. It is a useful field to investigate the role of representations in learning, because graphs are key tools for statistical reasoning (Section 1.3). Statistics education is also a suitable field for investigating the role of the computer in the classroom, because the computer is almost indispensable in performing genuine data analysis due to the large amount of data and laborious graphing (Section 1.4).

In some countries such as the United States of America and Australia, students already learn some statistics when they are about ten years old, explaining why most available research at the middle school age comes from these countries (ACE, 1991; NCTM, 1989, 2000). In the Netherlands, students first encounter descriptive statistics when they are about 13 years old, and hardly any Dutch research into statistics education with younger students has been carried out.

It is clear from the growing need for statistical literacy and the relatively small amount of research and experience with 10 to 14-year-old students that we need an instruction theory for early statistics education. In short, an instruction theory for a specific domain is a theory of how students can be supported in learning that domain. It entails knowledge about students' statistical intuitions, the key concepts of statistics, the type of reasoning and thinking that is possible for specific age groups, and supportive instructional activities embedded in a longitudinal learning trajectory.

The purpose of the present research is therefore

to contribute to an empirically grounded instruction theory for early statistics education.

When we write 'statistics', we mean descriptive statistics and exploratory data analysis, not inferential statistics. In Chapter 2, we set out our points of departure and summarize the research literature relevant to our study, in particular that of Cobb, McClain, and Gravemeijer, which leads to the research questions of the present study.

1.2 Design research

One way to develop an instruction theory is by conducting design research. The strength of design research (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003) or developmental research (Freudenthal, 1991; Gravemeijer, 1994, 1998) is that it can yield an instruction theory that is both theory-driven and empirically based. A design research cycle typically consists of three phases:

1 preparation and design;
2 teaching experiment;
3 retrospective analysis.

The methodology of design research is described in more detail in Chapter 3. In that chapter we also present an overview of the teaching experiments we carried out in Dutch seventh and eighth-grade classes (age 11-13). The preparation phase consists of a historical study (Chapter 4) and a so-called didactical phenomenology (Chapter 5). The teaching experiments in grade 7 are described in Chapters 6 and 7, and the teaching experiment in grade 8 in Chapter 9. Retrospective analyses are presented in Chapters 6 to 9.

1.3 Symbolizing

To investigate the role of graphical representations in learning statistics we use semiotics, the science of signs and meaning. Semiotically seen, a graph is a sign, and a sign is defined as something that stands for something else for someone (Peirce, NEM). In our context, the first 'something' is mostly an inscription on paper or computer screen and the second 'something' is a mental construction, ideally a mathematical or statistical object. By 'someone' we mostly mean a student, but it could also be 'the' community of statisticians. A symbol is a sign for which the representational relation is conventional or arbitrary, and not based on likeness for instance (Peirce, NEM). A diagram is a sign representing relations. In statistics education, we

are mainly interested in diagrams and symbols.
At the end of the nineteenth century, the non-fixed relationship of a sign and its object was introduced by the philosophy of language and has been widely accepted ever since. In particular, it is acknowledged that a sign is always interpreted as referring to something else within a social context. For a statistician, for example, a sketch similar to Figure 1.1 signifies a normal distribution, but a student who does not know this distribution as a statistical object may interpret it as an image of a mountain. This indicates a fundamental learning problem: symbols in mathematics and statistics refer to objects that students still need to construct. This problem can supposedly be overcome if students start with simple symbols and meanings they attribute to them and gradually develop more sophisticated symbols and meanings. It is assumed that the process of symbolizing (making, using, and adjusting symbols) and the process of constructing meaning of such symbols co-evolve (Meira, 1995). How this co-evolvement proceeds is the theme of Chapters 8 and 9.

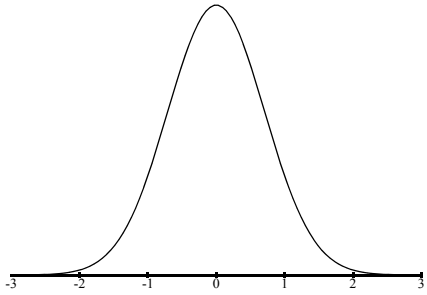


Figure 1.1: Graph symbolizing the normal distribution

1.4 Computer tools

Computer tools allow users to dynamically interact with large data sets and to explore different representations in a way that is impossible by hand. However, computer tools can also distract students' attention to the tools themselves instead of mediating effectively between the learner and what is to be learned (Noss & Hoyles, 1996). As we establish in Section 2.2, students with hardly any statistical background need special educational software to *learn* statistics. We are therefore interested in the question of how such educational computer tools could be used to support students' learning. In the present study we used the Minitools (Cobb, Gravemeijer, Bowers, & Doorman, 1997) that were designed for the Nashville research. The Minitools are three simple applets with one type of plot for each applet, which we have translated and revised. Minitool 1 offers a value-bar graph, Minitool 2 a stacked dot plot, and Minitool 3 a scatterplot (see Section 2.3). In the present study, students only used Minitools 1 and 2.

2 Background and research questions

The purpose of this research, as identified in Chapter 1, is to contribute to an empirically grounded instruction theory for statistics education at the middle school level. In this chapter, we begin with the pedagogical and didactical philosophy of the domain-specific instruction theory of Realistic Mathematics Education (RME). Next we survey the research in statistics education relevant to the present study and focus on the Minitools research from Nashville that the present study builds upon. In the last section we formulate the research questions of this study.

2.1 Realistic Mathematics Education (RME)

Realistic Mathematics Education (RME) is a theory of mathematics education that offers a pedagogical and didactical philosophy on mathematical learning and teaching as well as on designing instructional materials for mathematics education. RME emerged from research in mathematics education in the Netherlands in the 1970s and it has since been used and extended,[2] also in other countries. Some readers might wonder why we start with a theory on mathematics education as statistics is not a branch of mathematics. One reason is that contexts are very important to both the RME theory and statistics education. Moreover, educational practice is that statistics is taught as part of the mathematics curriculum.

The central principle of RME is that mathematics should always be meaningful to students. The term 'realistic' stresses that problem situations should be 'experientially real' for students. This does not necessarily mean that the problem situations are always encountered in daily life. Students can experience an abstract mathematical problem as real when the mathematics of that problem is meaningful to them. Freudenthal's (1991) ideal was that mathematical learning should be an enhancement of common sense. Students should be allowed and encouraged to invent their own strategies and ideas, and they should learn mathematics on their own authority. At the same time, this process should lead to particular end goals. This raises the question that underlies much of the RME-based research, namely that of how to support this process of engaging students in meaningful mathematical and statistical problem solving, and using students' contributions to reach certain end goals.

Views similar to those within RME have been formulated in general reform efforts in the United States (NCTM, 1989, 2000), Australia (ACE, 1991), and other countries, and by the theoretical movements such as situated cognition, discovery learning, and constructivism in its variants. The theory of RME, however, is especially

2. For instance: Freudenthal's work (e.g. 1973, 1983, 1991) and several dissertations (Treffers, 1987; De Lange, 1987; Van den Brink, 1989; Streefland, 1991; Gravemeijer, 1994; Van den Heuvel-Panhuizen, 1996; Menne, 2001; Van Amerom, 2002; Drijvers, 2003; Keijzer, 2003; Van den Boer, 2003).

tailored to mathematics education, because it includes specific tenets on and design heuristics for mathematics education. When we use the term 'design', we mean not only instructional materials, but also instructional setting, teacher behavior, interaction, and so on. RME tenets and heuristics are described in the following sections.

2.1.1 Five tenets of RME

On the basis of earlier projects in mathematics education, in particular the Wiskobas project, Treffers (1987) has defined five tenets for Realistic Mathematics Education:

1 *Phenomenological exploration.* A rich and meaningful context or phenomenon, concrete or abstract, should be explored to develop intuitive notions that can be the basis for concept formation.
2 *Using models and symbols for progressive mathematization.* The development from intuitive, informal, context-bound notions towards more formal mathematical concepts is a gradual process of progressive mathematization. A variety of models, schemes, diagrams, and symbols can support this process, provided these instruments are meaningful for the students and have the potential for generalization and abstraction.
3 *Using students' own constructions and productions*. It is assumed that what students make on their own is meaningful for them. Hence, using students' constructions and productions is promoted as an essential part of instruction.
4 *Interactivity.* Students' own contributions can then be used to compare and reflect on the merits of the different models or symbols. Students can learn from each other in small groups or in whole-class discussions.
5 *Intertwinement.* It is important to consider an instructional sequence in its relation to other domains. When doing statistics, what is the algebraic or scientific knowledge that students need? And within one domain, if we aim at understanding of distribution, which other statistical concepts are intertwined with it? Mathematics education should lead to useful integrated knowledge. This means, for instance, that theory and applications are not taught separately, but that theory is developed from solving problems.

In addition to these tenets, RME also offers heuristics or principles for *design* in mathematics education: guided reinvention, didactical phenomenology, and emergent models (Gravemeijer, 1994). We describe these in the following sections.

2.1.2 Guided reinvention

Freudenthal (1973, 1991) advocated teaching mathematics as a human activity as opposed to a ready-made system. When students progressively mathematize their own mathematical activity (Treffers, 1987) they can reinvent mathematics under the guidance of the teacher and the instructional design. This explains the first principle of RME, guided reinvention, which states that students should experience the learning of mathematics as a process similar to the process by which mathematics was

invented (Gravemeijer, 1994). The designer of realistic mathematics instruction can use different methods to design instruction that supports guided reinvention.
The first method is what Freudenthal called a 'thought experiment': designers should think of how they could have reinvented the mathematics at issue themselves. In fact, this is what Freudenthal (1991) used to do when he read mathematical theorems: find his own proof of the theorems. The second method is to study the history of the topic at issue: the method of carrying out a so-called historical phenomenology is used in Chapter 4. The third method, elaborated by Streefland (1991), is to use students' informal solution strategies as a source: how could teachers and designers support students' solutions in getting closer to the end goal?

2.1.3 Didactical phenomenology

To clarify his notion of phenomenology, Freudenthal (1983a) distinguished thought objects (*nooumena*) and phenomena (*phainomena*). Mathematical concepts and tools serve to organize phenomena, both from daily life and from mathematics itself. A phenomenology of a mathematical concept is an analysis of that concept in relation to the phenomena it organizes. This can be done in different ways, for example:

1 *Mathematical* phenomenology: the study of a mathematical concept in relation to the phenomena it organizes from a mathematical point of view. The arithmetical mean is used, for example, to reduce errors in astronomical observations. See Section 5.2 for a short mathematical phenomenology of distribution.
2 *Historical* phenomenology: the study of the historical development of a concept in relation to the phenomena that led to the genesis of that concept. For example, the mean evolved from many different contexts, including navigation, metallurgy, and astronomy. It took until the sixteenth century before the mean of two values was generalized to more than two values. The first implicit use was in estimating large numbers (Section 4.3.1).
3 *Didactical* phenomenology: the study of concepts in relation to phenomena with a didactical interest. The challenge is to find phenomena that "beg to be organised" by the concepts that are to be taught (Freudenthal, 1983a, p. 32). In Section 6.2 we describe how students organized a picture of elephants into a grid and used a so-called 'average box' to estimate the total number of elephants in the picture.

In this research, the design of instructional materials was preceded by a historical study of the relevant statistical concepts and graphs, including average values, distribution, and sampling. The goal of this historical study was to find problem situations or phenomena that could provide the basis for the development of the mathematical concepts or tools we wanted students to develop. Such problem situations could lead to paradigmatic solutions that are first specific for that situation, but can

be generalized to other problem situations. This last possibility is worked out under the heading of emergent models.

2.1.4 Emergent models

As the second tenet of RME about models for progressive mathematization implies, we search for models that can help students make progress from informal to more formal mathematical activity. Gravemeijer (1994, 1999a) describes how models *of* a certain situation can become a model *for* more formal reasoning. In the case of statistics, the notion of a distribution in combination with diagrams that display distributions was envisioned to become a model *of* data sets and later a model *for* more formal statistical reasoning (Gravemeijer, 2002).

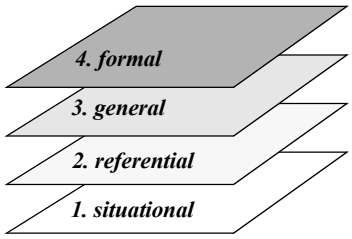


Figure 2.1: Levels of emergent modeling from situational to formal reasoning.

These four levels (Figure 2.1) can be described as follows (after Gravemeijer, Cobb, Bowers, & Whitenack, 2000, p. 243):

1 Situational level: activity in the task setting, in which interpretations and solutions depend on understanding of how to act in the setting (often in out-of-school settings);

2. Referential level: referential activity, in which models-*of* refer to activity in the setting described in instructional activities (mostly posed in school);

3. General level: general activity, in which models-*for* enable a focus on interpretations and solutions independently of situation-specific imagery;

4. Formal level: reasoning with conventional symbolizations, which is no longer dependent on the support of models-for mathematical activity.

2.2 Trends in statistics education research

Now that we have discussed the domain-specific theory of RME, we move to the domain of statistics education. This section gives an overview of research in statistics education relevant for the present study. Concentrating on research at the middle school level, when students are about 10 to 14 years old, we do not pay much attention to issues such as assessment (Gal & Garfield, 1997) and professional development of teachers (Mickelson & Heaton, in press; Makar & Confrey, in press), although these are important topics of research.

As background information we first sketch a short history of statistics education in the Netherlands. Statistics was proposed as part of the secondary mathematics curriculum in 1954, but the proposal was rejected due to concerns about an overloaded curriculum. It was not until the early 1970s that descriptive statistics was introduced in the upper-levels of the secondary mathematics curriculum. The main reasons for introducing statistics in the curriculum were that it was needed in many sciences and fields of work and that students could experience the usefulness of mathematics (Freudenthal, 1974; Van Hiele, 1974). Around 1990, statistics was introduced in the lower grades; since then graphs such as the box plot and the stem-and-leaf plot have been taught in grade 9 (W12-16, 1992).

We start our overview of research in statistics education with the work of Kahneman, Tversky, and their colleagues because much of the early research in statistics education was grounded in their work (Kahneman & Tversky, 1973, 1982; Shafir, Simonson, & Tversky, 1993; Tversky & Kahneman, 1971, 1982; Tversky & Shafir, 1992). From the early 1970s onwards, these cognitive psychologists have been investigating how people use statistical reasoning in everyday situations to arrive at decisions. Although these heuristics result in the same decisions as would be made based on statistical theory, there are instances when they lead to decisions which are at odds with such theory.[3] The overall impression is that statistical reasoning is very demanding cognitively.

In the late 1970s and 1980s, statistics education research mainly focused on college level and on probability (Pfannkuch & Wild, in press). In the 1990s, studies were undertaken at lower levels, mainly middle school level, and because statistics became part of the mathematics curriculum, more and more mathematics educators became involved in the discipline. Because statistics appeared early in the curriculum of the United States, Australia, and the United Kingdom, the majority of studies at the middle school level took place in these countries.

When we compare the curricula of different countries, for instance, the United States, the Netherlands, and Germany, we can observe a lot of variation. In the United States, mean, median, and mode are introduced in grade 4 or 5, when students are 9 or 10 years old (NCTM, 2000). In the Netherlands, students learn their first descriptive statistics in grade 8, when they are 13 or 14 years old. In most German states, the median is not even in the high school curriculum (Engel, personal communication, November 4, 2002).

These international differences in curricula indicate that we cannot simply apply research findings from, for instance, American studies to the Dutch situation. If Dutch students learn a statistics topic in a higher grade than American students, the Dutch students will mostly have a better mathematical background than American students in lower grades, which could make it easier (or perhaps more difficult) to learn cer-

3. For an overview in Dutch see Bakker (1999).

tain statistical topics. Mathematical issues that have been identified as important for statistics are multiplicative reasoning (ratios, percentages, and proportions), the Cartesian system, and line graphs (Cobb, 1999; Friel, Curcio, & Bright, 2001; Zawojewski & Shaughnessy, 1999). On the one hand, this means that Dutch students with proficiency in these matters do not necessarily encounter the same problems with statistical graphs and concepts as younger American students. On the other hand, the Dutch seventh-grade students do not have the same experience with data and science as most younger American students. This reveals the need for design research that will eventually lead to statistics curricula that are tuned to the mathematics and science curricula in the different countries.

The prominent image that emerges from reading the early research on statistics education is that students have trouble with understanding and using the mean, which was the most investigated statistic. Students mostly know how to calculate it but are not able to use it well (Hardiman, Well, & Pollatsek, 1984; Mevarech, 1983; Mokros & Russell, 1995; Pollatsek, Lima, & Well, 1981; Strauss & Bichler, 1988). Similar problems occur for many statistical graphs such as the histogram and for methods such as hypothesis testing. The literature gives the impression that students had to deal with artificial problems and artificial data (Singer & Willett, 1990), that statistics consisted of 'number crunching' (Shaughnessy, Garfield, & Greer, 1996, p. 209) and that statistics courses were overloaded with formal probability theory, which led to the nickname of 'sadistics' (Wilensky, 1997). With these images in mind, we summarize recent developments in statistics education research with the following five trends.

1 New pedagogy and new content;
2 Using exploratory data analysis (EDA);
3 Using technology;
4 Focusing on graphical representations;
5 Focusing on aggregate features and key concepts.

1. *New pedagogy and new content*. In the research literature, the information transmission model has made place for constructivist views, according to which students should be active learners (Moore, 1997). Bottom-up learning is advocated as opposed to top-down teaching. Students should have the opportunity to explore, build upon their intuitive knowledge, and learn in authentic situations. Such ideals are also expressed in the reform movements in mathematics education (NCTM, 2000) and the theory of Realistic Mathematics Education (2.1).

The old instructional paradigm was to teach the theory and then apply it. Freudenthal (1973, 1991) promoted the opposite: learning mathematical theory by developing it from meaningful situations that are usually seen as applications. Box (1999) advo-

cates a similar view for statistics education: statistics is often taught as a set of techniques, but these 'one-shot' standard ways of dealing with certain situations such as hypothesis testing are often black boxes to students, and do no justice to the practice of statistical investigations either. An example of the new pedagogy and content is given by De Lange and colleagues (1993), who report that students can reinvent meaningful graphs themselves.
A call for new pedagogy and content also came from a different source. Although statistics was (and still is) often taught as if it were a branch of mathematics, many statisticians have argued that it should be taught differently. G. Cobb (1997), for example, has spelled out the implications of this view for statistics education: where mathematics focuses on abstraction, statistics cannot do without context. One of his slogans to summarize this view was "more data and concepts, less theory and recipes" (see also G. Cobb & Moore, 1997). The emergence of constructivist philosophy and exploratory data analysis fitted well with these attempts to differentiate statistics from mathematics.

2. *Exploratory data analysis* (EDA) is a relatively new area of statistics, in which data are explored with graphing techniques (Tukey, 1977). The focus is on meaningful investigation of data sets with multiple representations and little probability theory or inferential statistics. In EDA, it is allowed to look for unanticipated patterns and trends in existing data sets, whereas traditional inferential statistics only allows testing of hypotheses that are formulated in advance.
EDA was adopted by several statistics educators to serve the need for more data and less theory and recipes (Biehler, 1982; Biehler & Steinbring, 1991). EDA was also thought useful to bridge the traditional gap between descriptive and inferential statistics. Descriptive statistics was often taught as too narrow a set of loosely related techniques, and inferential statistics tended to be overloaded with probability theory. EDA was considered an opportunity to actively involve students, to broaden the content of descriptive statistics, to give students a richer and more authentic experience in meaningful contexts, and to come closer to what statistics is really about (Ben-Zvi & Arcavi, 2001; Jones et al., 2001; Shaughnessy et al., 1996). Manually exploring real data with graphing techniques is laborious, which implies that using technology is a big advantage.

3. *Technology*, in particular computer software, enables to deal with larger data sets and allows for more efficient ways of representing data. It supplies ways of visualizing concepts, letting students interact with data sets, avoiding laborious computations, and allowing problem solving in real complex situations (Ben-Zvi & Friedlander, 1997). There is, however, an inherent tension between the new pedagogy and the use of technology. In the reform movement, students are seen as active creators

of knowledge and they should have the opportunity to experience mathematics as meaningful, but computer software mostly seems to provide the opposite: what is possible in a software program is predetermined, and what the computer does in reaction to clicking certain buttons can hide the conceptually important part of the operations (cf. Drijvers, 2003, about computer algebra systems). Of course, students need not know exactly which operations the software does to make a histogram, but they should be able to understand that data are categorized into certain intervals and that the bars' areas are relative to the number of values in those classes. In most statistical software packages such operations are hidden, which suggests that we need special software for education that minimizes this black-box character. As Biehler writes:

> Professional statistical systems are very complex and call for high cognitive entry cost. They are often not adequate for novices who need a tool that is designed from their bottom-up perspective of statistical novices and can develop in various ways into a full professional tool (not vice versa). (…) As a rule, current student versions of professional systems are no solution to this problem because they are technical reductions of the complete system. (Biehler, 1997, p. 169)

In Sections 2.3, we discuss bottom-up software tools that have been designed specially for middle school students and that have been used in this study, the Minitools (another bottom-up program, Tinkerplots (Konold & Miller, 2004), is under development). The present research focuses on cognitive tools (Lajoie & Derry, 1993), not on simulations, computer-based instruction, or web-based projects. We are interested in how educational tools can be used in a sensible way and ask what the graphs provided by these tools mean to students. These computer programs and graphs are tools that reorganize learning. We deliberately avoid 'amplify' or 'improve learning', because numerous researchers have shown that using computer tools often drastically changes the learning process and what is learned (e.g. Dörfler, 1993; Noss & Hoyles, 1996; Pea, 1987). It can be hard to compare statistical learning with and without the use of technology, because these two conditions lead to the learning of different 'content', which is, as noted above, often inseparable from the tools and symbols used (cf. Tall et al., 2000). In the present research, we ask *how* computer tools can be used, not *whether* using computer tools leads to better test results.

4. *Focus on graphical representations.* In EDA and technology, graphical representations play a central role. Graphs have the potential of showing more than the values in the data table. They can provide an overview of the whole data set and can highlight specific characteristics that are not visible from the numbers. Research has shown, however, that seemingly simple graphs are often not transparent to students (Ainley, 2000; Ainley, Nardi, & Pratt, 2000). It is acknowledged that transparency

is not a feature of the graph itself, but that it is connected to a purpose and tied to a cultural practice. Meira (1998) describes transparency as an emergent phenomenon intricately interwoven with learners' current activities and participation in ongoing cultural practices (see also Roth, 2003). This means that a graph is not an end goal in itself but that it should serve the purpose of solving a statistical problem and communicating the results. Even then, as Lehrer and Schauble (2001) note, graphs are too often taught as prefabricated solutions that have advantages and conventions that may be invisible to students. The question that rises is how can we let students reinvent and use graphical tools for statistical problem solving and in which order can we provide them with graphical tools that are meaningful for them.

Several researchers have proposed a certain order of graphs based on the complexity of the graphs for students (Biehler & Steinbring, 1991; Friel, Curcio, & Bright, 2001). Feldman, Konold, and Coulter (2000) suggest an order that is based on how easy it is to identify individual cases in a plot. Cobb, McClain, and Gravemeijer (2003) proposed a sequence of graphs, value-bar graphs before dot plots, which is summarized in the next section.

A conclusion we have drawn from several studies is that students should begin with graphs in which they can retrace each individual data value, so-called 'case-value plots', before they use graphs in which data are aggregated, so-called 'aggregate plots' (Konold & Higgins, 2003). Typical examples of case-value plots are case-value bar graphs (Figure 2.2) and stacked dot plots (Figure 2.3); examples of aggregate plots are histograms and box plots. Bar graphs can be both case-value plots and aggregate plots. If each bar represents a case, it is a case-value plot (a case-value bar graph, as Feldman et al., 2000, call it). If a bar represents a number of cases, a frequency for instance, it is an aggregate plot. A consequence of this is that using graphs such as histograms and box plots is more demanding for students than using bar graphs and dot plots. But there is more. There are hidden conventions and conceptual elements in histograms and box plots: in histograms, the area of the bars is relative to the number of values it signifies, and in box plots conceptual elements such as median and quartiles are depicted.

The difference in case-value and aggregate plots highlights that students need conceptual structures to conceive aggregates before they can interpret and use graphs sensibly. This insight explains the focus on aggregates and the key concepts of statistics at the middle school level (trend 5). In fact, this is the reason that we have a closer look at the relationship of graphs and concepts in the process of symbolizing and ask how we can support students in developing an aggregate view on data.

5. *Focus on aggregate features and key concepts*. Statistical data analysis is mainly about describing and predicting aggregate features of data sets.

> And so was born the modern sciences of statistics, the science of collections or aggregates, replacing the certainty of deterministic physics about a single individual by the certainty of the behaviour of groups of individuals. (Kendall, 1968, p. 6)

One can mostly not predict the values of individual cases, but one can predict aggregate features such as the mean or the distribution of a data set. Hancock, Kaput, and Goldsmith (1992) note that students have considerable trouble in conceiving data sets as aggregates. This implies that there is a strong conceptual component in interpreting graphs: in order to be able to see the aggregate students have to 'construct' it. After Hancock and colleagues, other researchers have dealt with the problem as well. Konold, Pollatsek, and Well (1997), for example, observed that high school students, after a yearlong statistics course, still had a tendency to focus on properties of individual cases, rather than on propensities of data sets. Ben-Zvi and Arcavi (2001) notice that seventh-grade students initially have a local view on data, and need time and good support to develop a global view on data. Without appropriate conceptual structures, students do not perceive the data as is necessary for doing statistical data analysis.

This stresses the necessity of addressing a number of key concepts of statistics, which are also called 'overarching statistical ideas' or, with a somewhat fashionable term, 'big ideas' (Schifter & Fosnot, 1993). There is no unique list of *the* key concepts of statistics. Here, we advocate a list for the middle school level that is inspired by a list from Garfield and Ben-Zvi (in press) for statistics education in general. We do not want to suggest any order, because these key concepts are only meaningful if dealt with in relation to each other (Wilensky, 1997).

- Variability;
- Sampling;
- Data;
- Distribution;
- Covariation.

Variability

Variability is at the heart of statistics because without variability there is no need for statistical investigation (Moore, 1990). In the PISA framework for assessment in mathematics education (OECD, 1999), statistics and probability fall under the overarching idea of 'uncertainty' (the others being 'change and growth', 'shape and space', and 'quantitative reasoning'). Kendall characterized statistics as the science that gives relative certainty about uncertainty.

> We live in a world of chance and change, surrounded by uncertainty; our object is to impose a purpose on that uncertainty and, in some sense, to bring the world under control (Kendall, 1968, p.1)

In fact, uncertainty and variability are closely related: because there is variability, we live in uncertainty, and because not everything is determined or certain, there is vari-

ability. If we plan to go somewhere by car, for example, it is uncertain at what time we will arrive due to variability among conditions. Various uncertain factors, such as traffic jams and weather conditions cause variability in travel times. We chose to call this key concept 'variability' because 'uncertainty' has the connotation of probability, which we do not address in this thesis.

Sampling
In order to investigate a variable phenomenon, one can take a sample. Without sampling there is no data set, distribution, or covariation to describe. The notion of sampling has been studied at the middle school level, but not extensively. Though students seem to have useful intuitions about sampling, it is a complex notion that is difficult to learn (Rubin, Bruce, & Tenney, 1990; Schwartz et al., 1998; Jacobs, 1999; Watson & Moritz, 2000; Watson, 2002).

Data
Sampling and measurement lead to data, that is numbers with a context (Moore, 1997). The idea of data involves insight into why data are needed and how they are created. This implies that the idea of data relies on knowledge about measurement. Too often, however, data are detached from the process of creating them (Wilensky, 1997), so many researchers and teachers advocate students collecting their own data (NCTM, 2000; Shaughnessy et al., 1996). This is time-consuming, so "talking through the process of data creation," as Cobb (1999) calls it as opposed to data collection, can sometimes substitute for the real creation of data. Once represented as icons, we can manipulate the data icons in a graph to find relations and characteristics that are not visible from the table of numbers (Lehrer & Romberg, 1996).
The idea of data is intimately connected with the other overarching ideas. For instance, data are often conceptually organized as 'pattern + deviation' (Moore, 1997) or as 'fit + residual' (G. Cobb, 1997) and data analysis is also described as the search for signals in noisy processes (Konold & Pollatsek, 2002). A data set can be analyzed for patterns and trends by using suitable diagrams. A key instrument in this analysis process is the concept of distribution with its various aspects such as center and spread.

Distribution
Distribution is probably the most central concept in statistics (Bethlehem & De Gooijer, 2000). It is an important tool in describing and predicting patterns in variability. Despite its complexity, it has recently been identified as an important key concept even for the middle school level (Cobb, 1999; Gravemeijer, 1999b, c; Petrosino, Lehrer, & Schauble, 2003). The problem is that students tend to see data sets as rows of individual numbers instead of a whole that can have aggregate features. Distribu-

tion could afford an organizing conceptual structure for thinking about patterns in variability and to view a data set as an aggregate. Besides that, distribution is a key idea in statistics that is connected to almost all other statistical concepts. Hence, focusing on distribution as an end goal of instruction or as a guideline for design has the additional advantage of possibly providing more coherence in the different statistical concepts and graphs (Cobb, 1999).
Because distribution is a complex concept with many aspects and different layers of possible understanding, from frequency distributions to probability density functions, research is needed to tease out the different conceptions of distribution that students have and can develop. The historical study in Chapter 4 aids in the analysis of this concept. Section 5.2 is devoted to this key concept of distribution, in particular the necessity of viewing a data set as a whole or an object that can have characteristics. A central question of the present study is how this process of becoming an object (reification) evolves (Chapter 8 and 9).
We consider center and spread as two main characteristics of the concept of distribution.

Center
Center can only be the center of something else, the distribution. If there is a 'true value', a measure of center can be used to estimate the signal in the noise (Konold & Pollatsek, 2002). However, as Zawojewski and Shaughnessy (2000) argue, students can only sensibly choose between the different measures of center if they have a notion of the distribution of the data.

Spread
Spread did not receive much attention in statistics education research until about 1997 (Shaughnessy, Watson, Moritz, & Reading, 1999; Reading & Shaughnessy, 2000). The reason for neglecting spread in education has probably been the focus on the measures of central tendency. We use the term spread for variation in the variable at issue, and 'variation' as a more general term, for example for variation in frequency, in density, or around a curve. Noss, Pozzi, and Hoyles (1999) report on nurses who view spread as crucial information when investigating blood pressure. Spread is also crucial information when doing repeated measurements, for example in a physical or biological context (Lehrer & Schauble, 2001). For an overview of the research on variation see Meletiou (2002).

Statistical covariation
Covariation is advocated as a topic for middle school by the NCTM (2000), but it is not in the Dutch mathematics curriculum of secondary schools. As a study by Cobb, McClain, & Gravemeijer (2003) in grade 8 shows, covariation is a difficult statistical

idea. They are convinced that students should first develop a notion of univariate distributions before they can sensibly deal with covariation. It is perhaps not too difficult to draw a straight line by eyeballing, but Cobb and colleagues chose to deal with covariation at a deeper level than just drawing a line through a cloud to signify a trend.
Covariation is not a topic investigated in the present study. We focused on the first four key concepts, because the students hardly had any statistical background and had only 10 to 15 lessons on data analysis (it was not possible to use more lessons). We do mention covariation as a key concept, because we think that it should be somewhere in the Dutch mathematics curriculum: students encounter covariation in physics, chemistry, biology, and the social sciences, mostly represented in scatterplots.

There is not much research on how these key concepts can be developed at the middle school level. If we narrow our focus to studies that use bottom-up software in which students do not have to choose from ready-made graphs, the only one research we can directly build upon is that of the research team in Nashville (e.g. Cobb, McClain, & Gravemeijer, 2003). That research focused on the notion of distribution (grade 7) and covariation (grade 8), and, as mentioned in Chapter 1, formed the starting point of the present study.
In the next section, we summarize the results of the Nashville research in a very detailed way, which allows us to present elements of our initial conjectured instruction theory for early statistics education and to confirm, reject, or extend their findings.

2.3 Nashville research with the Minitools

In this section,[4] we summarize the instructional design issues of the research carried out by Cobb, McClain, and Gravemeijer and their team at Vanderbilt University, Nashville, USA. We focus on their points of departure (2.3.1), the rationale of the software (2.3.2), results and recommendations (2.3.3), and the way they described the process of symbolizing (2.3.4). The numerous articles on these teaching experiments mainly deal with themes that transcend the instructional design issues that we are interested in here: individual and collective development of mathematical learning, sociomathematical norms and mathematical practices, tool use, the teacher's proactive role, diversity, equity, and so on. We therefore base this section mainly on Gravemeijer's manuscripts on the instructional design part of the research (Gravemeijer, 1999b, c, 2000a, b, 2001) and thus make the instructional design component accessible to a wider audience. Other sources used for this section were videotapes of the seventh-grade teaching experiment and articles that have been published (Cobb, 1999, 2002; Cobb & Hodge, 2002a, 2002b; Cobb & McClain, 2002, in press;

4. This section has been authorized by Cobb and Gravemeijer.

Cobb, McClain, & Gravemeijer, 2003; Cobb & Tzou, 2000; Gravemeijer, 2001, 2002; McGatha, 1999; McClain, 2002; McClain & Cobb, 2001; McClain, McGatha, & Hodge, 2000; McGatha, Cobb, & McClain, 2002; Sfard, 2000a). We start with factual information about the teaching experiments with the Minitools.
Members of the research team at the Vanderbilt University, Nashville, TN, USA, were Paul Cobb, Kay McClain, Koeno Gravemeijer, Maggie McGatha, José Cortina, Lynn Hodge, Carrie Tzou, Carla Richards, and Nora Shuart-Farris. External consultants were Cliff Konold (University of Massachusetts, Amherst) and Erna Yackel (Purdue University, Calumet). A short pilot with seventh-graders (twelve years old) was carried out in the fall of 1996 (McGatha et al., 2002). The seventh-grade teaching experiment took place in a regular school situation in the fall of 1997 and consisted of 37 lessons of 40 minutes, with a group of 29 students over a ten-week period; the eighth-grade teaching experiment was one year later in 1998 with 11 of the 29 students (thirteen years old) and consisted of 41 lessons of 40 minutes over a fourteen-week period. This last experiment was carried out during the so-called 'activity hour', normally used for doing homework. The school served a mixed working class and middle class student population, and about 40% of the students were classified as minority (almost all African American). In the course of the experiment, 5 students dropped out and 11 voluntarily continued to attend throughout (Cobb, personal communication, November 22, 2002).

2.3.1 Points of departure and end goal

The team's general points of departure can be described with respect to the trends of the previous section. In terms of pedagogy (trend 1), the theory of RME served as a domain-specific instruction theory, and heuristics such as the guided reinvention principle and emergent models were used for instructional design (Section 2.1). In line with Freudenthal's credo that students should experience mathematics as a human activity, the team strived for students' participation in genuine data analysis as opposed to learning a set of techniques. As pointed out in Section 2.2, EDA (trend 2) has an investigative spirit that seemed suitable for this purpose. EDA can be done without difficult probabilistic concepts and, in comparison to traditional descriptive statistics, it forms a broader basis for authentic learning in meaningful contexts. Graphical software can be supportive of the goals of EDA (trend 3) but, as pointed out earlier, special software was needed to allow students to build upon what they already knew (case-value plots) and that at the same time could lead to more sophisticated graphical representations (aggregate plots). Because there was no software that suited the team's needs, it had to be designed especially for this project. To avoid students experiencing statistics as just "doing something with numbers" (McGatha et al., 2002) and to stay within the investigative spirit of EDA, graphs were a central focus as instruments in problem solving (trend 4). However, as pointed out in the paragraphs on the 'key concepts', students also need to develop statistical con-

cepts when they learn to analyze data (trend 5). Otherwise, they would probably keep perceiving individual data instead of conceiving the aggregates displayed in the graphs.
We have numbered the issues to refer more easily to the same issues later in this book when we show how our results compare to those of the Nashville team. P# stands for point of departure number #, and R# for result or recommendation number #. We would like to draw attention to the fact that the first teaching experiment took place in 1997, when much less was known about statistics education than nowadays, and that issues in this section should be seen as part of a design *process*.

P1. As pointed out, EDA requires conceptual tools to conceive of patterns and trends. The research team decided to combine, as they called it, the *process character* of EDA with the *product character* of statistical tools and concepts as endpoints of instruction (Gravemeijer, 2000a; McGatha et al., 2002).

P2. The pedagogical point of departure was that the learning process should always be genuine data analysis from the outset, as opposed to learning specific tools and then applying them. Students should deal with real problems that they considered *socially important* or, in other words, a statistical problem had to be solved *for a reason*. This point of departure influenced the contexts that were used (AIDS, ambulance travel times, CO_2 emission, speeding cars, braking distances of cars).

P3. Because data collection is time-consuming and not always possible, the research team decided "*to talk the students through the process of data creation,*" as they called it. Too often, in their view, data are just collected for the purpose of doing something with data and graphs. The term 'creation' was deliberately used to stress that data are not already there, but need to be produced (G. Cobb & Moore, 1997; Lehrer & Romberg, 1996; Moore, 1997; Roth, 1996). Data are the result of a process of measurement that usually involves several decisions. First, there must be a question that implies a certain way of looking at a phenomenon. It has to be decided which attribute is to be measured and how this is to be done. For some types of measurement, this procedure is clear-cut, for instance for height. For other contexts, this is not directly evident, for instance, measuring whether a certain medicine has the desired effect. The team found it important that students always considered the whole process of data creation. To achieve this, typically the teacher during whole-class discussion capitalized on questions such as, "What is it that we want to know and why do we want to know this? How do we come to know it?" After students had thought about the data creation process, the teacher told them how the data at issue had been created: "This is how they did it". The students then got the opportunity to comment on this.

P4. Instructional format. The instructional format typically used was a whole-class discussion, followed by individual or group work with computers, and again a whole-class discussion. The first discussion was for talking through the data creation process and the second discussion was for collectively discussing students' analyses.

P5. Distribution. With these points of departure, the question was, "Which concepts and graphical tools do we have to focus on?" In her dissertation on the learning of the research team, McGatha (1999) describes how the notion of *distribution* as a potential end goal of the instructional sequence emerged from a conceptual analysis of the topics and from the discussion on the relevant research literature. The two major reasons to focus on distribution were the following.

First, distribution is a concept that can help students to conceive aggregates. Data had to be seen as distributed in a space of possible values and the notion of distribution had to become, in the end, an object-like entity with characteristics such as where the data were centered and how they were spread. Conventional graphs such as box plots and histograms had to emerge as ways to describe and structure distributions. A more extensive analysis of this notion of distribution is given in Section 5.2.

Second, focusing on distribution could bring more coherence to the curriculum. Traditionally, statistical concepts and graphs are taught in middle school as a set of loosely related techniques (Cobb, 1999). Concepts such as mean, median, spread, and skewness are only meaningful when they refer to distributions. Coherent statistical knowledge could emerge if these concepts can be developed as characteristics of distributions and graphs as ways to structure and describe distributions.

P6. The team conjectured that *comparing distributions* would be a valuable way to involve students in reasoning about characteristics of distribution, because there would be a reason to think about the whole data sets.[5] Hence, most instructional activities in the seventh-grade experiment involved comparing two distributions.

P7. Mean. There were three reasons not to focus on the *mean*. The first reason was that a considerable amount of research, including an analysis of the pilot (McGatha et al., 2002), showed that students generally see the mean as a procedure and do not use it sensibly to compare groups (see also Section 2.2). The team thought that this procedural approach would interfere with the kind of reasoning they wanted to foster. Gravemeijer (1999b):

> Another reason for avoiding the mean was in our assertion that the mean is not very relevant if it comes to characterizing the distribution of data values. The mean does

5. This idea is in line with findings of Watson & Moritz (1999) and Konold & Higgins (2002).

> give you an aggregated measure of the data values but it does not tell you anything about the distribution as such. (p. 13)

As we explain in Chapters 4 and 5, this is not our own view.

P8. The *median* seemed more appropriate to use as a measure of center than the mean, and quartiles seemed more appropriate to use as a measure of spread than the standard deviation (Gravemeijer, 1999b). In the team's reasoning, the median was easier to understand than the mean (but see R16), the position of the median in relation to the extreme values showed skewness, and a five-number summary of minimum, first quartile, median, third quartile, and maximum was seen as a way to describe the distribution (Tukey, 1977).

P9. Sampling. EDA typically does not deal with probability and sampling issues. G. Cobb and Moore (1997), writing about college level, recommend EDA as a basis for data production (design), which in turn forms the basis for probability and then statistical inference. Because the research team brought distribution and multiplicative reasoning to the foreground, the issue of *sampling* moved to the background and was not to be addressed explicitly.

P10. The designers experienced that it was almost impossible to find real data sets that served their pedagogical agenda. This was the reason they mainly used *realistic* data as opposed to real data, where realistic means "as it could have been in a real situation."

P11. Multiplicative reasoning. During the pilot, the team had noticed that students struggled with ratios and proportions when solving statistical problems. Students' reasoning is called additive if they use absolute instead of relative frequencies. For instance, if students compare 4 of a group of 24 with 6 of a group of 65 without taking the proportion into account, they reason additively. If they use proportions or percentages to compare the groups of different size, they reason multiplicatively (cf. Thompson, 1994). Konold et al. (1997) argue that a focus on a proportion of data within a range of values is at the heart of a statistical perspective, so multiplicative reasoning is an important issue to address. Multiplicative reasoning "also cuts across a number of content strands and constitutes the overarching goal for American mathematics instruction at the middle-school level" (McClain & Cobb, 2001, p. 108). The team therefore decided to make multiplicative reasoning a central theme of the research.

> Our goal for the learning of the classroom community was that reasoning about the distribution of data in multiplicative terms would become an established mathematical practice that was beyond justification. (Cobb, 2002, p.176)

The problem of technology as described in trend 3 implied that special software had to be designed. This is the topic of the next section, which also deals with the representational backbone of the instructional sequence. The specially designed so-called 'Statistical Minitools' (Cobb, Gravemeijer, Bowers, & Doorman, 1997) were used in 27 of the 34 classroom sessions of the seventh-grade experiment, and in about 25 of the 41 sessions of the eighth-grade experiment.

2.3.2 Rationale of the Statistical Minitools[6]

Most educational statistical software packages only provide common and culturally accepted graphical representations (e.g. Datascope, Fathom, VU-Stat; see also Biehler, 1997). Users of such packages need to anticipate the information that they can extract from the representations and measures that the computer program has to offer. For students with hardly any statistical background, this is difficult, and such an approach is not in line with the reform movements or the RME philosophy. Tabletop (Hancock, 1995) does offer non-standard plots, but in the team's view this program seemed to orient students to focus on characteristics of individual cases, which was at odds with their concern for distribution. The team therefore developed three so-called Statistical Minitools to support a process of guided reinvention of the notion of distribution and graphical representations that can organize distributions. The backbone of the sequence is formed by a series of inscriptions that are embedded in the Minitools.

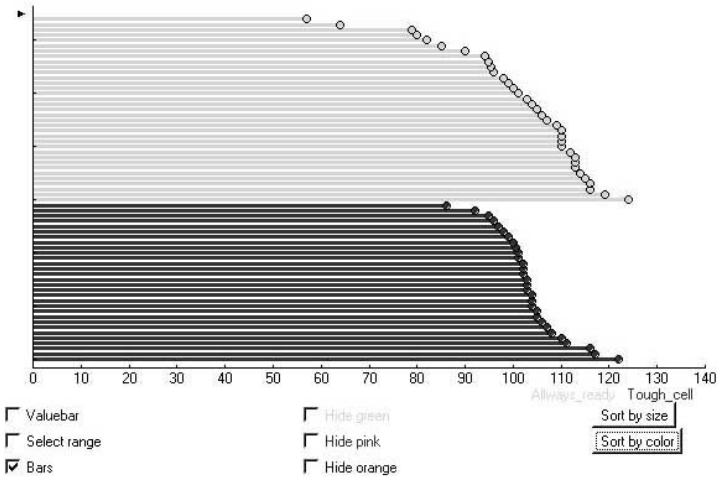


Figure 2.2: Case-value bars in Minitool 1 signifying life spans in hours of two different battery brands

6. This section is an edited version of a section by Gravemeijer (2001). In this way, we stick as closely as possible to the original motivation for the Minitools design. The Minitools are free for use in the classroom (NCTM, 2003; www.wisweb.nl).

The rationale is that the activities with the computer tools follow each other in such a manner that the activity with the newer tool is experienced as a natural extension of the activity with the earlier tool.

Minitool 1

The starting point is in the measures, or magnitudes, that constitute the data set. With Minitool 1, case-value bars (Figure 2.2) are introduced that signify single measures. Initially, the measures under investigation are of a linear type, such as length and time. Later, this is generalized to other types of measures. Students can organize the data by sorting by size and by color (for comparing subsets). Furthermore, there are two tools, a 'value tool' and a 'range tool', which have to support the organization of data and understanding of the variable on the horizontal axis. The value tool (or 'reference line') was also meant to afford visual estimation of a measure of center, in particular to avoid calculations for getting the mean. The range tool was to support the discourse on a modal class.

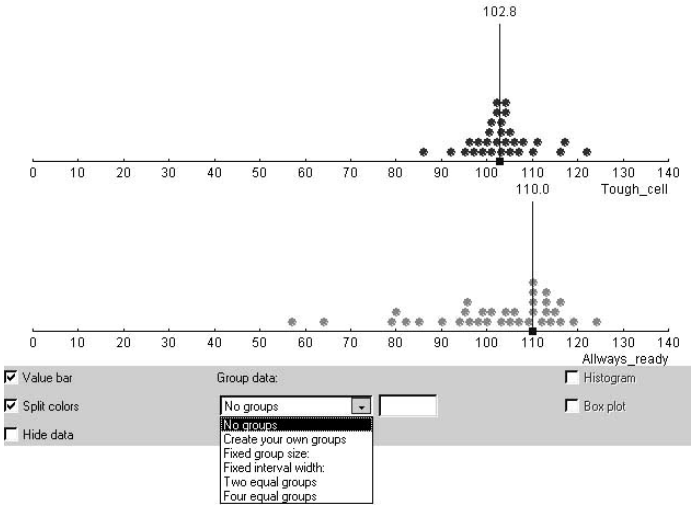


Figure 2.3: A dot plot in Minitool 2

Minitool 2

In the discussions on data represented by value bars, the students are assumed to focus on the endpoints of the value bars. As a consequence, these endpoints come to signify the corresponding value bars for the students. This allows for the introduction of a dot plot as a more condense inscription, that omits the value bars and only keeps the endpoints (Figure 2.3). The dots are collapsed down onto the axis without moving to the right or left. This explains the open spaces that sometimes occur in a

dot plot. Note that there is no frequency axis: the students still need to construct one mentally. This type of representation is also called a numberline plot.[7]
Minitool 2 offers various tool options to help the students structure the distribution of data points in a dot plot. There are five options to structure the data: making your own groups, making two or four groups with an equal number of data points, making groups of a certain size (e.g. 10 data points per group), and making equal intervals.

Create your own groups. Apart from incorporating options that would anticipate conventional graphs, the team wanted to make sure that the Minitool would offer the students the freedom to structure the data in ways that make sense to them. Hancock and colleagues had observed that students tend to make their own clusters of data (Hancock, Kaput, & Goldsmith, 1992), so an option to create one's own groups was built in.

Two and four equal groups. Minitool 2 features equal-groups options, one that partitions a set of data points into two halves, and one that partitions a set of data points into four quartiles. The latter results in a four-equal-groups inscription, which can be seen as a precursor to the conventional box plot. The similarity with a box plot is clearer when the 'hide data' option is used (Figure 2.4).

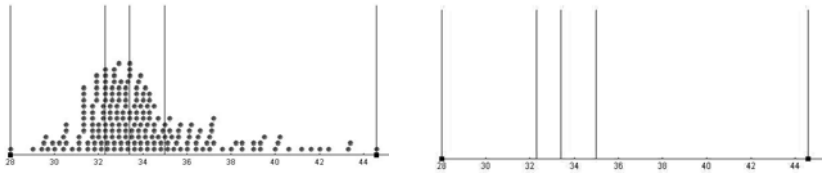


Figure 2.4: Dot plot structured into four equal groups and with 'hide data'

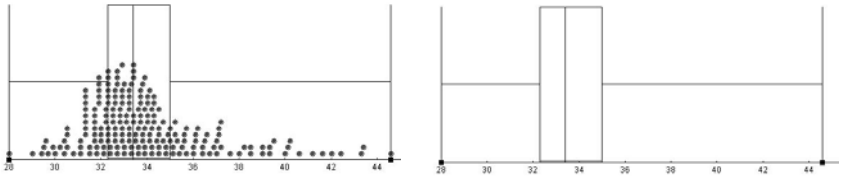


Figure 2.5: Box plot overlay that was made available in the revised version of Minitool 2

7. Some people find the stacked dot plots unclear because the vertical dimension only gives an informal sense of density. This problem could be solved by rounding off values or changing the scale such that the dots stack neatly, so that the vertical dimension conveys frequency.

The hide data option also creates the opportunity to shuttle back and forth between the dot plot and the four-equal-groups inscription. This is helpful, for instance, if one wants to make the adequacy of a four-equal-groups description a topic for discussion. The four-equal-groups option was considered useful for fostering multiplicative reasoning: it does not depend on the size of the data set, so data sets of different size can be compared adequately. The team assumed that the four-equal-groups inscription (a precursor to the box plot) would support reasoning about the shape of the distribution in terms of density, which is a key characteristic of a distribution. The idea is that students learn to relate the width of an interval to the density of the distribution in that area.

Equal interval width. As an alternative to the four-equal-groups option, the students can also choose to structure the data in equal intervals. When using this option, the students can indicate the size of the interval width. In addition to intervals, Minitool 2 shows the number of data points in each interval (Figure 2.6).

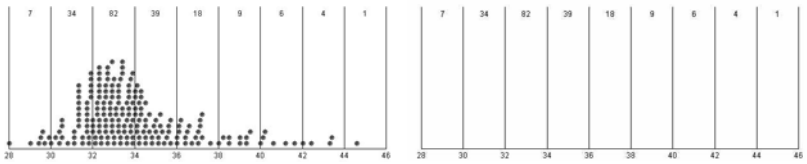


Figure 2.6: Dot plot structured into equal intervals and with ‘hide data’

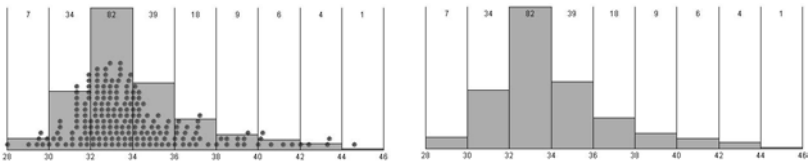


Figure 2.7: The histogram overlay that was made available in the revised version of Minitool 2

The hide data option makes it possible to view just the numbers. The corresponding inscription can be seen as anticipating a frequency histogram. In a more elaborate sequence, the histogram with absolute frequencies might be developed into a histogram with relative frequencies. Such a histogram might eventually come to function as a basis for the description of a distribution by a continuous curve: the curve might be the result of taking the limit with the interval width approaching zero. However, the notion of taking a limit is far too demanding for seventh-graders. Instead, the end goal was formulated as follows: “The notion of the shape of the distribution has to be the result of analyzing data sets with the Minitools, after students have come to

see the shape of the dot plot as a description of the distribution of variable values in terms of density" (Gravemeijer, 2000b). The box plot and the histogram overlays (Figures 2.5 and 2.7) were not available in the original version of Minitool 2; they were added during the present research, but only the eighth graders worked with them.

Minitool 3

For the eighth-grade experiment in Nashville, the end goal was that students would come to see bivariate data as a series of distributions of univariate data. To support a learning process towards that end goal, Minitool 3 was developed. It offered a scatterplot with a few ways of structuring the data such as groups of a fixed size or equal groups. One insight was that a series of stacked distributions (as in Figure 2.8) would support deeper understanding of covariation in a scatterplot (Cobb, McClain, & Gravemeijer, 2003). Deeper understanding meant that covariation had to be much more than drawing a line through a cloud of dots. In the team's view, covariation in a scatterplot can be seen as a description of how a univariate distribution changes along with the independent variable. This underlines the importance of a good understanding of univariate distributions before addressing statistical covariation. For this transition students need to see a series of dot plots turned at angles and squeezed into one vertical line.

Because Minitool 3 was not used in the present study, we do not discuss this Minitool any further (see for instance www.wisweb.nl). We only discuss the findings when students worked with Minitool 3 that were relevant for our own study.

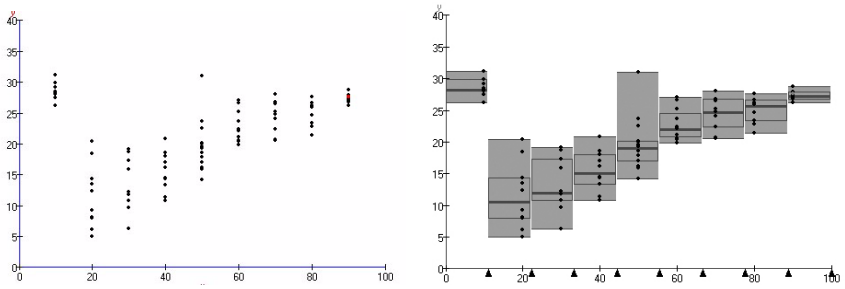


Figure 2.8: Scatterplot in Minitool 3 with stacked data as opposed to a 'clouds' (left) and four equal groups within the scatterplot slices (right)

2.3.3 Results and recommendations

In this section, we summarize the results and recommendations that formed starting points of the present study.

Seventh-grade teaching experiment

R1. Cobb and Tzou (2000) claim that *talking through the data creation process* proved successful. One sign of that is that students adopted the teacher's questioning attitude on the data creation.

R2. It was noted in retrospect that talking through the data creation process is a way of implicitly addressing the issue of *sampling*. However, it was acknowledged that sampling should be addressed more thoroughly, since it underlies the notion of fair comparison and it cannot always be kept in the background (Gravemeijer, 1999c). This issue is revisited in R17.

R3. The students used the *range tool* to partion data sets and to isolate intervals to make comparisons as well as to investigate consistency. For example, in the context of the life span of two battery brands (Always Ready and Tough Cell), Casey used it to indicate the interval 104.9 to 126.6 (Figure 2.9):

Casey: Alright, because there's ten of the Always Ready and there's ten of the Tough Cell, there's 20, and half of 20 is ten.

Teacher: And why would it be helpful for us to know about the top ten, why did you choose that, why did you choose ten instead of twelve?

Casey: Because I was trying to go with the half. (Cobb, 2002, p.178)

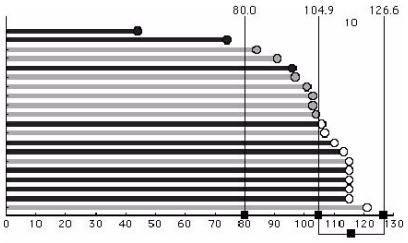


Figure 2.9: Battery data set in Minitool 1 with range tool

R4. The *value tool* or vertical reference line was used to find certain values or to reason with certain cutting points. For example, Brad used it at 80 (Figure 2.9):

Brad: See, there's still green ones [Always Ready] behind 80, but all of the Tough Cell is above 80. I would rather have a consistent battery that I know will get me over 80 hours than one that you just try to guess.

Teacher: Why were you picking 80?

Brad: Because most of the Tough Cell batteries are all over 80. (Cobb, 2002, p. 178)

R5. There turned out to be a shift from looking at individual cases to the data set as

a whole. “The mathematical practice that emerged as the students used the first minitool can (...) be described as that of exploring qualitative characteristics of collections of data points” (Cobb, 2002, p. 179). The term ‘collection of data points’ is used here in opposition to ‘distribution’. The first term refers to additive reasoning whereas the second denotes multiplicative reasoning. The notion of the majority of data did not become a topic of discussion until students analyzed data sets with unequal numbers (Cobb, 2002) and the notion of majority was then used in conjunction with the hill notion (see next item).

R6. In the context of comparing situations before and after the installment of a speed trap, one student spontaneously used the notion of a ‘hill’ to refer to the majority of the data points. The meaning attributed to the hill, was more than just visual; it was interpreted in the context of the speed of cars (Figure 2.10):

> Janice: If you look at the graphs and look at them like hills, then for the before group the speeds are spread out and more than 55, and if you look at the after graph, then more people are bunched up close to the speed limit which means that the majority of the people slowed down close to the speed limit. (Cobb, 2002, p. 180)

In describing hills, Janice had reasoned about qualitative relative frequencies. This notion of the majority of the data, however, did not become a topic of discussion until the students analyzed data sets with unequal numbers of data values.

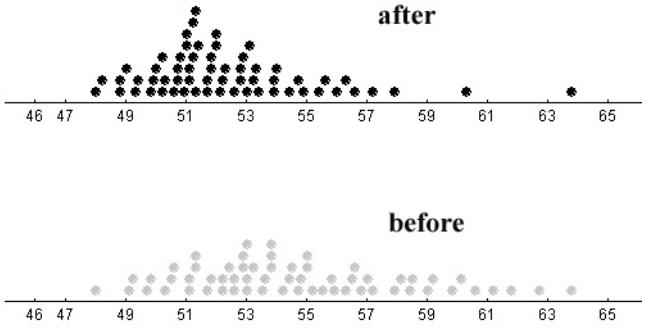


Figure 2.10: Speed trap data set in Minitool 2 (in miles per hour).

The team’s analysis indicates that a majority of the students could interpret graphs of unequal data sets organized into equal interval widths, an analog of histograms, and into four equal groups, an analog of box plots, in terms of global characteristics of distributions. The mathematical practice that emerged as they developed these competencies can be described as that of exploring qualitative characteristics of dis-

tributions (Cobb, 2002). Note that this second mathematical practice is about distributions, whereas the first practice is about collections of data points.

R7. Bottom-up approach. The team has been reasonably successful in the bottom-up approach. The first example that supports this claim is that students themselves came up with the notion of consistency to denote what we would perhaps call the variation or spread of the data. The second example is students' spontaneous use of the notion of a 'hill' for comparing distributions (see student quotes in R6).

R8. Multiplicative reasoning versus distribution. Gravemeijer (2000b) remarks that when students use Minitool 2 grouping options, there is a pitfall of fragmentation, because students often compared vertical slices. For instance, students noticed that 50% of one data set was below a certain cutting point whereas 75% of the other data set was below the same point. Hence, the development of multiplicative and arithmetical reasoning had some success at the expense of developing the notion of distribution as a whole in relation to shape. See also R13.

R9. Students did not use the value tool to estimate means, except for one student, but he used the value tool only as an instrument of calculation (Gravemeijer, personal communication, January, 2003).

R10. Though students started to talk about a shift of the data, for instance in the context of a speed trap, students found it very difficult to describe the shift in *quantitative* terms. Presumably, students did not conceive the measures of center as characteristics of the whole distribution.

R11. For the instructional sequence, the team considered it sufficient if students developed an image of the normal distribution as symmetrical, with a maximum in the middle, and descending towards the extremes (Gravemeijer, 2000b).

R12. As a way of letting students reason about the appropriateness of graphical displays, the team had gradually come to push the students in the role of *data analysts* as opposed to just problem solvers. By that the team meant that students did not need to solve the problem themselves, but should describe and represent the data in such a way that others, such as politicians and other decision makers, could make a reasonable decision.

Eighth-grade experiment on covariation

Because covariation is included in the American middle school mathematics curriculum and is important for statistical literacy, the team felt that bivariate data and

scatterplots had to be addressed. During the eighth-grade experiment they learned to see bivariate data as a distribution of univariate distributions. For students, it turned out to be very difficult to achieve this view, but the intermediate step of using stacked data that show a series of vertical distributions was promising (Cobb, McClain, & Gravemeijer, 2003).
In the Dutch mathematics curriculum up to grade 12, covariation and scatterplots are not dealt with. Because we did not address covariation in the present study (apart from the spontaneous reinvention of a scatterplot in Section 6.11), we only summarize results from the eight-grade experiment that are essential for the present research.

R13. In many situations, for example when working with data sets with a lot of variation and especially when working with univariate data, students found it hard to perceive hills. Cobb and colleagues (2003) conclude that *smoother* distributions are probably needed to support students' perception of shapes. It is also advisable to foster a basic *set of shapes* such as normal, skewed, and bimodal.

R14. The research team used the term 'shape' for more than the visual part. It refers to how the density of the data varies over the measured variable. Students talk about full and empty parts: data are bunched up in certain parts of the graph or are spread out in other parts, for instance in a four-equal-group organization. In that sense, they indeed started to reason about density, and not just about frequency distributions.

R15. The team concluded that the *mean* cannot really be avoided (see *P7*). The main reason that Gravemeijer (1999b) mentions is that many real data values are already means:

> It may be noted that the eighth-grade teaching experiment on bivariate data made us very aware of the fact that many interesting data sets are constituted by data that consist of means themselves. For instance, when looking at the relation between heart diseases and alcohol consumption, based on data from various countries. We might use the very same type of situations, where the question of how to compensate for the differences in population size of the various countries could be a starting point. (p. 13)

This last idea is worked out by Cortina and colleagues (Cortina, 2002; Cortina, Saldanha, & Thompson, 1999).

R16. The *median* turned out to be problematic, both for the students and for the instructional designers. For the students, the median initially was a procedure of counting inward from each end of an ordered data set to find the middle number. The team defined the median as the value that divides the data set in half (with the option of two equal groups in Minitool 2) and tried to let students use the median as an indi-

cation of the majority, but were not successful in doing so. There was only one episode in which a student said that the median is where the majority is, but this idea was rejected by the other students. Cobb and colleagues (2003) write that students saw the median at the level of just another data point instead of a measure of center. With hindsight, the team realized that the median had previously been used as a cutting point to support multiplicative reasoning about equal parts of the data (for example, 25% versus 75%) and not as a measure of center (see also R8).

R17. Sampling. Around the 36th lesson, the team made an interesting observation. Students did not want to predict a single value (reaction time of an eighth-grader), but argued that the reaction times of another group of ten eighth-graders would probably be similar. Gravemeijer writes that the team had 'downplayed' the role of sampling. They had kept the fact that most problems dealt with samples in the background, for they wanted to avoid problems with probability. The few times they tried something with making inferences about single events proved them right, but the example above "opened a new avenue." "We can address sampling, if the questions are cast in terms of the population from which the sample is taken" as opposed to comparing data sets (Gravemeijer, 1999b, p.15-16). Gravemeijer goes on to say:

> To foster the idea that there is a strong relation between the distribution of a population and the distribution of an adequate sample of that population, students have to gain experience with comparing samples and populations. (ibid.)

These seventeen results were the main issues that the present research built upon. We now turn to the issue of symbolizing, which was another focus of both the Nashville and the present research.

2.3.4 Symbolizing

As mentioned in Chapter 1, the purpose of the present research is to contribute to an instruction theory for early statistics education. In Section 2.2 we hinted at the relationship between graphs and concepts: it is impossible to make sense of graphs without having appropriate conceptual structures, and it is impossible to communicate about concepts without any representations. Thus, to develop an instruction theory it is necessary to investigate the relation between the development of meaning of graphs and concepts. This process is a focus of both the Nashville research and the present research, and part of that process is called 'symbolizing'. What exactly is meant by symbolizing? Before answering that question, we invite the reader to try and think of the normal distribution without any representation or application of it. We have no idea of how to do that. In our minds, we see the bell shape or perhaps another graph; we think of the probability density function, the Galton board, or we think of phenomena that can be modeled with the normal distribution (for example, height). In line with Dörfler's observation that he could not find the concept of the

number 5 or the triangle in his mind, we cannot find the concept of the normal distribution in our mind, only representations (Tall, Thomas, Davis, Gray, & Simpson, 2000). As Peirce (CP 2.228),[8] one of the founding fathers of semiotics, remarked: all our mathematical thinking is on signs. In particular, graphs and diagrams are integral to statistical reasoning. But how do graphs become meaningful for students? How do diagrams come to serve as useful tools to organize data and solve statistical problems? What is the relation between the concept and a graph of the normal distribution?

In recent years, researchers in mathematics education have framed such questions as semiotic questions and have taken graphs as signs or symbols. A sign, in short, is something that stands for something (referent or object) for someone. Many concrete entities can serve as a sign: graphs, diagrams, charts, tables, icons on a computer screen, sketches on paper, building blocks, or a knot in a handkerchief. A symbol is a special type of sign, namely one that is arbitrary or conventional in a sense. While a footprint or a photograph is mostly not used as a symbol, the letter π is commonly used in mathematics as a symbol for the proportion of circumference and diameter of a circle. It has generally been acknowledged since at least the end of the nineteenth century (Frege, 1962) that there is no fixed relation between a sign and a referent. Someone has to interpret a sign in relation to a referent. But how does a person learn how to interpret a sign in the intended way?

It is now commonly accepted in the philosophy of language (and mathematics education research) that symbols gain their meaning by the way they are used (e.g. Wittgenstein, 1984), that is in activity and in a community of practice (Wenger, 1998; Meira, 1998). A histogram, for instance, has no meaning by itself, but only when used to solve a certain problem or to describe the distribution of a data set. Thus it can function as a sign of a distribution.

Literally, symbolizing means "making a symbol." The term 'symbolizing' as it is used by the research community, however, seems to be a *pars pro toto*: it stands for the whole process of making a symbol for a specific purpose, using it, improving it and possibly making a new symbol (Cobb, Yackel, & McClain, 2000; Gravemeijer, Lehrer, Van Oers, & Verschaffel, 2002). The term is used to stress that symbols gain meaning in activity. This attention for symbolizing as an activity as opposed to symbols as ready-made representations is in line with Freudenthal's (1991) focus on mathematics as the activity of mathematizing versus mathematics as a ready-made system.

In mainstream psychology, internal and external representations are distinguished: signs are external representations and concepts internal representations. It is however not clear how a connection is established between internal and external represen-

8. Following common practice we refer to the Collected Papers of Peirce as CP with the volume (2) and section number (228).

tations. According to Cobb (2000), it is by the focus on the activity of symbolizing that the dichotomy between internal and external representations can be overcome. Several proposals have been made to describe this process of symbolizing. Latour (1990), Meira (1995) and Roth (1996) write about 'cascades of inscriptions'; Lacan (1968), and in his footsteps many others, used a 'chain of signification' (Cobb, 2002; Gravemeijer et al., 1997; Presmeg, 2002; Walkerdine, 1988; Whitson, 1997).
For the origin of the chain of signification, we have to go back to another founding father of semiotics, Ferdinand de Saussure, who conceived a sign as a pair of signifier and signified: the signifier signifies the signified. Lacan used this idea to describe how the meaning of one sign can slide under another sign; in this way a chain of signification can be built.

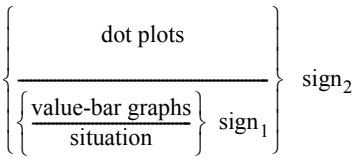


Figure 2.11: Chain of signification with the Minitools after Cobb (2002, p.188).

Cobb (2002) uses this idea to describe how the meaning of the value-bar graph in Minitool 1 slid under the next sign, the dot plot of Minitool 2 (Figure 2.11). The chain of signification started at the problem situation of life span of batteries and other contexts. After talking through the process of data creation (P3), the life spans were inscribed as horizontal bars in Minitool 1. This inscription (used as a sign) opened possibilities to partition the data set, for example, and to discuss features of the data set (extreme values, intervals, range). The first mathematical practice (R5) is defined as reasoning about qualitative characteristics of collections of data points. Students routinely investigated the number of data points above or below a certain value or within a specific interval. This habit was continued when students worked with Minitool 2. The way in which they used this computer tool to partition data sets was consistent with and built upon the ways of organizing data they were used to with the first Minitool. Cobb (2002) notes that it is the way of reasoning with a sign that can slide under a new sign, not something that emerges from the material features of the signs themselves.
While reasoning with the second Minitool, students came to view data sets as distributions as opposed to collections of individual data points. Cobb (2002, p. 187) claims that "the central mathematical idea of distribution would not have emerged had the students continued to use the first minitool or if the second minitool had involved a different way of describing data." And "it is doubtful that the line plot inscription [dot plot of Minitool 2] would have afforded the emergence of the notion

of distribution had the students used the second minitool from the very beginning of the design experiment" (ibid.). He represents the chain of signification with the Minitools as in Figure 2.11.
In our view, there are several problems associated with using chains of significations to analyze students' learning, for example if two chains come together when comparing signs. We have therefore searched for an alternative semiotic framework without that linear nature. We discuss those issues in Chapters 8 and 9.

2.3.5 Conclusions for the present study

The conclusions of the Nashville research with the Minitools that were most relevant for the instructional design part of the present study were the following:

- *Center*. The measures of center had to be reconsidered. It turned out that the mean should be addressed and that problems around the median had to be solved. How do these statistics become measures of the data set or distribution?
- *Shape of distributions*. More attention should be paid to the shape of distributions in terms of density (shift of the hill; where are data bunched up?). The pitfall of fragmentation due to a focus on multiplicative reasoning should be avoided (R8). Using smoother unimodal distributions without comparing numbers of data points in certain groups might help to avoid the problems mentioned in R8 such as that students do not consider the whole shape when comparing parts multiplicatively.
- *Sampling* has to be taken into account from an early stage onwards. The notion of distribution could perhaps be developed by using sampling (R17), but how this can be done without turning to sampling distribution is an open question.
- *Symbolizing*. According to Cobb (2002) students' learning process with the two Minitools can be described with a chain of signification and mathematical practices. In Chapter 8 we argue why we have searched for a different semiotic framework and chose for Peirce's semiotics.

2.4 Research questions

Because the present study is a sequel to the design research just described, we took the same points of departure, the same Minitools, the results and recommendations of these experiments as the basis for the present research. All of the numbered issues of Section 2.3 are addressed in the following chapters to show in which sense our results confirm the Nashville results or deviate from them. Here we only highlight the key issues to motivate our research questions.
In the Nashville research, the end goal was distribution as a single, multifaceted notion (P4). As indicated in the section on key concepts (2.2), we aim for simultaneously dealing with data, distribution, and sampling in a coherent way, with distribution as the central idea. The first research question of this thesis, to be understood within the RME approach, is similar to the one of the Nashville research:

1. How can students with little statistical background develop a notion of distribution?

Because of the issues on sampling, we decided from the outset to pay more attention to sampling than the Nashville team did (R17). The answer to this first question is given in Chapter 7.
The Nashville team claims that there is a reflexive relationship between how students use symbols and what they signify for students (e.g. Cobb, 2002). Cobb uses the notion of a chain of signification to describe this process, but we argue in Chapter 8 there are practical and theoretical problems with the theory of chains of signification. Although the Nashville team was able to describe the development of mathematical practices on the macro-level with chains of signification, it is not exactly clear to us how this signification process evolves at the micro-level. To contribute to an instruction theory for early statistics education, we found it necessary (2.2 and 2.3.4) to investigate the symbolizing process, the relationship between graphical and conceptual development, in more detail. The second main research question is therefore:

2. How does the process of symbolizing evolve when students learn to reason about distribution?

We use different semiotic theories as instruments of analysis in our search for a theory that helps to describe and analyze this process of symbolizing, and that can inform instructional design as well. We answer the second research question in Chapter 8. Because the two research questions turn out to be strongly related when using Peirce's semiotics, an integrated research question is answered in Chapter 9.

3 Methodology and subjects

If you want to understand something you have to change it
and if you want to change something you have to understand it.
Seth Chaiklin (personal communication, June 17, 2002)

There is nothing so practical as a good theory.
Kurt Lewin (1951, p. 169)

3.1 Design research methodology

The purpose of the present research is to contribute to an empirically grounded instruction theory for early statistics education. Such a theory should specify patterns in students' learning as well as the means supporting that learning in the domain of statistics education (Cobb, Confrey, et al., 2003). This implies that the development of an instruction theory includes both the design of such instructional means and research of how these means support successive patterns in students' reasoning. Particularly if the research aims at specific types of learning that differ from common educational practice, one needs to design instructional materials that support the desired type of learning. In general, we first need to create the conditions in which we can develop and test an instruction theory, but to create those conditions we also need research. Design and research are therefore highly intertwined when developing an instruction theory.

In the present research we are especially interested in how students can learn to reason about distribution in RME-oriented education. This implies that we need to design an instructional environment that supports such learning and that we need to anticipate successive patterns in students' reasoning that could lead to particular end goals.

Our methodology falls under the general heading of design research, because it considers design as a crucial part of the research. This type of research is also termed 'developmental research', because instructional materials are developed, but by using the term 'design research' we hope to avoid two possible connotations (cf. Cobb, 1999). One is developmental psychology in the style of Piaget, and the other is research that describes the development of mathematical concepts in students. Design research (Edelson, 2002; Kelly & Lesh, 2000), developmental research (Engeström, 1987; Gravemeijer, 1994; Van den Akker, 1999), and design experiments (Brown, 1992; Collins, 1992) all treat design as a strategy for developing and refining theories. These types of research have been used successfully in a wide range of domains and for a variety of research questions (Edelson, 2002; Educational Researcher 32(1)). Design experiments can also be combined with comparative empirical research (Brown, 1992).

Cobb, Confrey, et al. (2003) identify five features that apply to different types of design research. The first is that its purpose is to develop theories about learning and

the means that are designed to support that learning. We develop an instruction theory for early statistics education and instructional means that support the learning of a multifaceted notion of distribution. The second feature of design research is its interventionist nature. The methodology allows researchers to take their "best bets" (Lehrer & Schauble, 2001) at all times so that they are not constrained to improve the design after an experiment cycle has been carried out. The third cross-cutting feature is that design research has a prospective and reflective component that need not be separated by an experiment. In implementing hypothesized learning (the prospective part) the researcher confronts conjectures with actual learning that he observes (reflective part). The fourth feature is the cyclic character of design research: invention and revision form an iterative process. Conjectures on learning are sometimes refuted and alternative conjectures can be generated and tested. The fifth crosscutting feature of design research is that the theory under development has to do real work (see the motto of Kurt Lewin: Nothing is so practical as a good theory). This theory is relatively humble in the sense that it is developed for a specific domain, for instance statistics education. Yet it must be general enough to be applicable in different contexts such as classrooms in other schools in other countries.

The objectives of design research are different from those of comparative empirical research. The main objective of design research is to develop theories together with instructional materials whereas the main objective of comparative research to evaluate theories or materials. This does not mean that we separate developing and evaluating theories, because in design research the theory that is under development is evaluated during and after design experiments. Glaser and Strauss' (1967) remarks about their method of comparative analysis also apply to design research: "Although our emphasis is on generating theory rather than verifying it, we take special pains not to divorce those two activities, both necessary to the scientific enterprise." (p. viii)

The difference in emphasis of the objectives also implies different norms of justification of theories. The Research Advisory Committee of the National Council of Teachers of Mathematics (RAC, 1996) observes a gradual shift from norms that apply to comparative empirical research to norms that apply to research such as design research. The norm of justification in the first type is often limited to assessing *whether* innovative curricula or professional development programs are better than traditional ones. The norm of justification in the second type is that an empirically grounded theory about *how* the design works, in terms of anticipatory conjectures, can be tested and revised in practice.

Design research is evaluated against the metrics of innovation and usefulness, and its strength comes from its explanatory power and grounding in experience. Moreover, it often leads to products that are useful in educational practice because they have been developed in practice. Design research, as we use it, consists of cycles of three phases:

1 a preparation and design phase,
2 a teaching experiment,
3 a retrospective analysis.

The results of such a retrospective analysis mostly feed a new design phase. In this way, the retrospective analysis of the Nashville research formed the basis for the present design research. The design research of the Nashville team had resulted in developed software, a set of instructional activities and an emerging instruction theory for early statistics education aiming at the notion of distribution as a single multifaceted notion. This did not mean that we could simply replicate this research because the Dutch context differs considerably from the American. The exploratory interviews indicate, for instance, that Dutch students in grade 7 have a different mathematical and statistical background than those in Nashville. To answer the research questions of this study, we needed to develop instructional activities that would be suitable within the Dutch context while taking into account new insights from the prior experiments, such as the role of the mean, median, and sampling. We also wanted to enhance the software, for instance by allowing histograms and box plots in the interface. Additionally, there were still open questions around the process of symbolizing. These could only be answered from a situation in which students would get the opportunity to share their ideas and make their own graphs. Before discussing the three phases of a design research cycle in more detail we need to define what we mean by a hypothetical learning trajectory.

3.2 Hypothetical learning trajectory (HLT)

A design and research instrument that proved useful during all phases of design research is the so-called 'hypothetical learning trajectory' (HLT), which we regard as an elaboration of Freudenthal's thought experiment. Simon (1995) defined the HLT as follows:

> The hypothetical learning trajectory is made up of three components: the learning goal that defines the direction, the learning activities, and the hypothetical learning process–a prediction of how the students' thinking and understanding will evolve in the context of the learning activities. (p. 136)

Simon used the HLT as part of the so-called Mathematics Teaching Cycle, mostly for one or two lessons, but we use it as an instrument in design research for longer sequences of instruction.[9]

The HLT is the link between an instruction theory and a concrete teaching experiment. It is informed by general domain-specific and conjectured instruction theories

9. The TAL team uses so-called 'Learning-Teaching Trajectories', but these refer to longitudinal curriculum strands across several grades (Van den Heuvel-Panhuizen, 2001). Our HLT is more similar to what Klaassen (1995) calls a scenario.

(cf. Gravemeijer, 1994), and it informs researchers and teachers how to carry out a particular teaching experiment. After the teaching experiment, it guides the retrospective analysis, and the interplay between the HLT and empirical results forms the basis for theory development. This means that an HLT, after it has been mapped out, has different functions depending on the phase of the design research and continually develops through the different phases. It can even change during a teaching experiment.

1 During the design phase, the HLT, once formulated, guides the design of instructional materials that have to be developed or adapted. The confrontation of a general rationale with concrete activities often leads to a more specific HLT, which means that the HLT usually develops during the design phase (Drijvers, 2003).
2 During the teaching experiment, the HLT functions as a guideline for the teacher and researcher what to focus on in teaching, interviewing, and observing. It can happen that the teacher or researcher feels the need to adjust the HLT or instructional activity for the next lesson. As Freudenthal wrote (1991, p. 159), the cyclic alternation of research and development can be more efficient the shorter the cycle is. Minor changes in the HLT are usually made because of incidents in the classroom such as anticipations that have not come true, strategies that have not been foreseen, activities that were too difficult, and so on. In such cases, a micro-cycle of design, experiment, and analysis occurs within a macro-cycle of design research. Such micro-cycles are generally not accepted in comparative empirical research, but in this type of design research, changes in the HLT are made to create optimal conditions and are regarded as elements of the data corpus. This means that these changes have to be reported well and the information is stronger if changes are supported by theoretical considerations. The HLT can thus also change during the teaching experiment phase.
3 During the retrospective analysis, the HLT functions as a guideline determining what the researcher should focus on in the analysis. Because predictions are made about students' learning, the researcher can contrast those anticipations with the observations made during the teaching experiment. Such an analysis of the interplay between the evolving HLT and empirical observations forms the basis for developing an instruction theory. After the retrospective analysis, the HLT can be reformulated, in an often more drastic way than during the teaching experiment, and the new HLT can guide a next design phase.

An HLT can be seen as a concretization of an evolving instruction theory. Conversely, the instruction theory is informed by evolving HLTs. For example, if patterns of an HLT stabilize after a few macro-cycles, these generalized patterns in learning or instruction and the insights of how these patterns are supported by instructional means can become part of the emerging instruction theory.

For an overview of the levels in the design research as conceived in our case, see Table 3.1; these levels are meant to be neither completely exclusive nor exhaustive. For some readers, the term 'trajectory' might have a linear connotation. Although we aim for a certain direction, the HLT was non-linear in the sense that we did not make a linear sequence of activities in advance that we strictly adhered to (cf. Fosnot & Dolk, 2001).

Table 3.1: Levels in the present design research

theories and knowledge from multiple sources	psychology, educational science, semiotics, activity theory, mathematics, mathematics education, statistics, statistics education, science education,...
instruction theories	ranging from small to large domain, e.g. from statistics at the middle school to Realistic Mathematics Education in general
hypothetical learning trajectories	particular teaching experiments
instructional materials (means of support)	activities, software, tools, teacher guides, etc.

Between lessons we often took a slightly different next step than thought before and sometimes even adjusted the end goals (micro-cycles of design research). Moreover, in our use of the term 'learning trajectory', we do not mean to exclude the teaching component of education. A better term might be 'education trajectory'.
In the following sections we give a more detailed description of the three phases of a macro-cycle of design research and discuss relevant methodological issues.

3.3 Phase 1: Preparation and design

It is evident that the relevant present knowledge about a topic such as statistics education should be studied first. Gravemeijer (1994) characterizes the design researcher as a tinkerer or, in French, a *bricoleur*, who uses all the material that is at hand, including theoretical insights and practical experience with teaching and designing. In our case, relevant theoretical knowledge came from multiple sources such as mathematics, statistics, realistic mathematics education, history, psychology, general educational research, cultural-historical activity theory, philosophy of language, linguistics, and semiotics.
In the first design phase, we collected and invented a collection of activities that could be useful and discussed these with colleagues who were experienced in designing for mathematics education. We also adjusted activities that had proven successful in Nashville. The important criterion for selecting an activity was its potential role in the HLT towards the end goal of distribution. Would it possibly lead to types of reasoning that students could build upon towards that end goal? Would it be

challenging? Would it be a meaningful context for students?
Our first HLT included expectations about students' learning with the activities and Minitools used in the Nashville research and was mainly informed by their points of departures and results (Section 2.3). For instance, we assumed that students would initially view data points as individual data without looking at the whole structure of the data set. We further assumed that it was important to talk through the process of data creation and let students adopt a role as data analyst. Another source of inspiration was the history of statistical concepts (Chapter 4).

3.4 Phase 2: Teaching experiment

The notion of a teaching experiment arose in the 1970s. Its primary purpose was to experience students' learning and reasoning first-hand, and it thus served the purpose of eliminating the separation between the practice of research and the practice of teaching (Steffe & Thompson, 2000). Over time, teaching experiments proved useful for a broader purpose, namely as part of design research. During a teaching experiment, researchers and teachers take their 'best bets', as Lehrer and Schauble call it (2001). That is, they use activities and types of instruction that seem most appropriate at that moment according to the HLT. Observations in one lesson and theoretical arguments from multiple sources can influence what is done in the next lesson. Hence, this type of research is different from experimental research designs in which a limited number of variables is manipulated and effects on other variables are measured. The situation investigated here, the learning of students in a new context with new tools and new end goals, is simply too complicated for such a set-up. Besides that, a different type of knowledge is looked for, as pointed out earlier in this chapter: we do not want to assess innovative material or a theory, but we need prototypical instructional materials that could be tested and revised by teachers and researchers, and an instruction theory that can be used by others to formulate their own HLTs suiting local contingencies.
For a careful retrospective analysis, it is necessary to keep track of changes in the HLT and of students' learning. The data collection during the teaching experiments varied (see Table 3.2). In the last two seventh-grade experiments and the eighth-grade experiment, we collected student work, tests before and after instruction, field notes, audio recordings of whole-class discussions and mini-interviews, and video recordings of every lesson and, in grade 8, of the final interviews. During the first teaching experiment, we spontaneously interviewed students during the lessons to get to know what graphs meant to them and how they had solved the problems. Soon afterwards, we came to call these interventions 'mini-interviews', lasting from about twenty seconds to four minutes, and looked for ways to systematize these mini-interviews. During the first teaching experiments, we had namely noted that some students were easier to interview than others, and that some were more interesting or

explicit in their utterances, which resulted in non-representative samples of the class. In the second experiment, we decided to focus on four students with average learning abilities to get a better image of the majority of the students.

Table 3.2: Chronological overview of the teaching experiments and data collection. The levels are *mavo* (lower general secondary education), *havo* (higher general secondary education), *vwo* (pre-university education).

Class (# students)	type of experiment	data collection	# of lessons	level
26 students grade 7	interviews (15 minutes per pair)	audio recording	-	*mavo, havo, vwo*
1A (25) grade 7	4 exploratory lessons	audio, student work, field observations, mini-interviews	4	*havo*
1F (27)	1st teaching experiment	idem plus final test	12	*vwo*
1E (28)	2nd teaching experiment		15	*vwo*
1D (23)	not attended	only teacher notes	10	*havo-vwo*
1C (23)	3rd teaching exp	audio, video, student work, observations, mini-interviews, pre-test and final test	12	*havo*
1B (23)	4th teaching exp (two assistants)		12	*havo*
18 classes (by June 2003)	implementation at one school	field observations in five lessons, e-mail conversation with two teachers, some student work	about 200	*havo, havo-vwo, and vwo*
2B (30) grade 8	5th teaching exp (three assistants)	audio, video, student work, observations, mini-interviews, final test, and interviews between lessons and after last lesson	10	*havo-vwo*

In the last seventh-grade experiment and the eighth-grade experiment, when assistants helped us, we divided the class into three groups and interviewed all students in our own group with approximately the same frequency. We formulated questions beforehand that we would ask the students and all interviewers had to ask at least these questions. These questions were motivated by the anticipations of our HLT and by specific questions we were interested in (e.g., do students prefer a certain Minitool to investigate the spread of the data?). In this way, we tried to get a representative image of the classes' developments. These mini-interviews were an important source of information for the evolving HLT. We realize that these mini-inter-

views had a learning effect, since they mostly evoked reflection. The validity of the research was not in danger in our eyes, because the question was not whether the designed instructional sequence was suitable for other classes and teachers as well, but (1) how students could learn certain notions and graphs in a process of guided reinvention and (2) how the process of symbolizing evolved. Moreover, most mini-interview questions were formulated in advance and discussed with assistants. This means these questions can be viewed as part of the HLT.

During the last seventh-grade experiment, the assistants were two students in mathematics who specialized in mathematics education, and during the eighth-grade experiment there were three pre-service teachers of mathematics of whom two were always present in the classroom. We prepared them before the teaching experiments and audio taped briefings after almost every lesson.

Our teaching experiments all took place in regular classrooms in a regular school situation. In this study, the teacher and the researcher were not the same person, except for one lesson in class 1B (grade 7) when the teacher could not be there. To avoid the risk of extra 'noise' caused by inexperienced teachers, we worked with two teachers with 26 and 11 years of experience, who were also used to participating in teaching experiments.

The sequence designed for the seventh grade was later used by two novice teachers in eighteen other classes (by June 2003). Occasional visits and written reports from the teachers did give some indication of the effect of having experienced teachers and one or more researchers conducting mini-interviews. This issue is taken up in Chapter 10.

Apart from these lessons given by the novice teachers, we attended all lessons of the teaching experiments. As the remarks about the mini-interviews earlier in this section make clear, the researchers were not just observers because they influenced the learning process. In design research, that is even the objective. The knowledge we need includes the ways in which we deal with unexpected situations to foster the learning process desired. As a participant in the learning culture, it is easier for the researcher to experience the factors relevant to the HLT or instruction theory and which were not foreseen explicitly (De Corte, 2000). Such factors can range from issues on the micro-scale (e.g. poor knowledge of percentages for instance or the usefulness of a computer projector), via the meso-scale (e.g. school culture: students in one experiment were not used to class discussions) to issues on the macro-scale (e.g. the Dutch mathematics curriculum).

3.5 Phase 3: Retrospective analysis

In the retrospective analysis the HLT is compared with students' actual learning. On the basis of such analyses we can answer the research questions and contribute to an instruction theory. For the last teaching experiment in grade 7 and the one in grade 8, we used the following method of analysis.

First we transcribed the episodes that could inform us about the topics of interest. What constituted a topic of interest was determined by the research questions and the HLT. Off-task behavior ("Can I have my pen back?"; "My mouse is not working") was not transcribed, but coded as such (e.g. 'soc' for social talk and 'comp' for computer problems). Then we used a method that is inspired by the 'constant comparative method' (Glaser & Strauss, 1967; Strauss & Corbin, 1998) and Cobb and Whitenack's method of longitudinal analyses (1996). We read all transcripts and watched the videotapes chronologically episode-by-episode. With the HLT and research questions as guidelines, conjectures about students' learning and views were generated, documented, and tested at the other episodes and other data material (student work, field notes, tests). This testing meant looking for confirmation and counter-examples. The process of conjecture generating and testing was repeated, not on the level of the conjectures themselves, as Cobb and Whitenack (1996) did, but on the original data material. Seemingly crucial episodes were discussed with colleagues to test whether they agreed upon our interpretation or perhaps could think of alternative interpretations (peer examination).

For the analysis of the last seventh-grade experiment we used computer software for coding the transcripts, namely MEPA, which stands for 'multiple episode protocol analysis' (Erkens, 2001). We coded students' utterances line-by-line with task (e.g. battery, jeans), literal terms (e.g. mean, spread out, Minitool), the concept dealt with (e.g. spread, distribution), and clues for easy retracing (e.g. 'see video'). These codes proved useful during the retrospective analysis to retrace all instances of a certain kind. An example of a conjecture tested in this way was that students find it easier to estimate the mean in Minitool 1 than in Minitool 2 (see Chapter 7 and the Appendix). In both classes 1B and 2B, the number of transcript lines was about 10,000.

For the eighth-grade experiment, all transcripts were coded with conjectures that were related to the HLT and the research questions. The transcripts of the lessons which seemed most informative (4, 6, 7) with the conjectures attached to episodes were discussed with the three assistants. Only codes that all four of us agreed upon were kept as codes. In Chapter 9 we provide examples of such conjectures and one example of a conjecture that one of us found not so clear in that episode. In this way, about a quarter of the transcripts was discussed. There were very few codes that any of us found doubtful, from which we concluded that the agreement among judges was high. The results of the retrospective analysis formed the basis for adjusting the HLT and for answering the research questions.

3.6 Reliability and validity

A few methodological issues have already been discussed in previous sections. Here we deal more explicitly with *reliability* as the absence of unsystematic bias and *validity* as the absence of systematic bias (Maso & Smaling, 1998). The issues discussed in this section are inspired by guidelines of Maso and Smaling (1998) and Miles and Huberman (1994).

Internal reliability refers to the reliability within a research project. It can be improved with several methods. In the previous sections, we discussed the data collection, how we coded the transcripts, how we used computer software during one retrospective analysis, and how we discussed the coding with three assistants. Internal reliability further refers to the reasonableness and argumentative power of inferences and assertions. We improved that by discussing the critical episodes, including those discussed in this book, with colleagues (peer examination).

External reliability usually denotes replicability, meaning that the conclusions of the study should depend on the subjects and conditions, and not on the researcher. In qualitative research, replicability is mostly interpreted as virtual replicability; the research must be documented in such a way that it is clear how the research has been carried out and how conclusions have been drawn from the data. A criterion for virtual replicability is 'trackability' (Gravemeijer & Cobb, 2001; Maso & Smaling, 1998). This means that the reader must be able to track the learning process of the researchers and to reconstruct their study: failures and successes, procedures followed, the conceptual framework used, and the reasons to make certain choices must all be reported. In Freudenthal's words:

> Developmental research means: experiencing the cyclic process of development and research so consciously, and reporting on it so candidly that it justifies itself, and that this experience can be transmitted to others to become like their own experience. (1991, p. 161)

Internal validity refers to the quality of the data collections and the soundness of the reasoning that has led to the conclusions (also labeled as 'credibility'). We used several methods to improve the internal validity of this study.

- During the retrospective analysis, we tested conjectures that were generated and tested at specific episodes at other episodes and other data material, such as field notes, tests, and other student work (source triangulation). During this testing stage we searched for counterexamples of our conjectures.
- The succession of different teaching experiments made it possible to test the conjectures developed in earlier experiments in later experiments.
- We analyzed important episodes with multiple theoretical instruments of analysis (theoretical triangulation). See for instance Section 8.1.

– Theoretical claims are substantiated where possible with transcripts to provide a rich and meaningful context. The possibility to do this in longer texts is the major reason to write this thesis as a book and not as a collection of journal articles.

External validity is mostly interpreted as the generalizability of the results. The question is how we can generalize the results from these specific contexts as to be useful for other contexts. An important way to do so is by framing issues as instances of something more general (Cobb, Confrey, et al., 2003; Gravemeijer & Cobb, 2001). The challenge is to present the results (instruction theory, HLT, instructional activities) in such a way that others can adjust them to their local contingencies (Barab & Kirshner, 2002). In the conclusions we list a number of issues of the Nashville team that have been confirmed in the present study (10.3). Such issues have become more general. Additionally, we found patterns that occurred in several classes of our own teaching experiments (Chapters 6, 7, 9, Appendix).
By using the general theory of semiotics, we intend to provide more abstract explanations connected to a theoretical network that is beyond the immediate study (theoretical validity). To investigate the role of graphs and their meaning for students, we use semiotics to frame graphs (in relation to object and interpretant) as signs and the learning process as signification (Chapters 8 and 9).
In addition to generalizability as a criterion for external validity we mention 'transferability' (Maso & Smaling, 1998). If lessons learned in one experiment are successfully applied in other experiments, this is a sign of successful generalization. Gravemeijer and Cobb (2001), for instance, claim that the method of analysis they had developed in a teaching experiment on measurement in grade 1 proved useful in the statistics experiments described in Section 2.3. This implies that the transferability and viability of the results of the present study can better be judged in the future if applied in other situations.

3.7 Overview of the teaching experiments and subjects

3.7.1 Exploratory interviews

We first wanted to know to what extent the activities and results of the Nashville experiments would apply to the Dutch situation, because we expected differences between the Dutch and American students. On the one hand Dutch students quickly learn to calculate their own grades for their reports with means from grade 7 onwards (unlike American students). On the other hand, Dutch students tend to learn fewer statistical graphs in earlier grades (mainly bar graphs). It is difficult to test students' prior knowledge of distribution. We were therefore focused on their prior knowledge of statistics, and we were especially interested in Dutch students' notion of the mean and their understanding of the graphs in Minitool 1 and 2. To that purpose, we inter-

viewed 26 students randomly chosen from the different seventh-grade classes of a school in Amsterdam, which offered three levels of education: *mavo* (the lower general secondary education track), *havo* (the higher general secondary education track), *vwo* (the pre-university track). At that time, about 15% of the Dutch students attended *vwo*, 20% *havo*, and 35% attended *mavo*. About 50% of the students of that school belonged to an ethnic minority. The students were interviewed in pairs for about 15 minutes. The data collection included audio recordings and student work. For the results of these interviews see Section 5.1.

3.7.2 Grade 7 experiments

The aim of these experiments was to find ways to teach the notion of distribution in relation to other key concepts of statistics and the types of graphs that structure and display distributions. The teaching experiments in grade 7, when students are 12 or 13 years old, were held at a secondary school in a small city near Utrecht in the school year of 1999-2000. This *havo-vwo* school had about 800 students from grade 7 to 12. The school is considered a 'white' (ethnically mainly Dutch) school serving a middle-class population. The teacher had been teaching for 26 years, and had considerable experience as a curriculum designer. Each week we discussed the next two lessons. Every odd-numbered lesson was in a regular classroom without computer or a computer projector. Every even-numbered lesson took place in a computer lab where the students worked in pairs with Minitools 1 and 2. These teaching experiments were carried out in what is called 'theme education': students study a theme for about five or six weeks (12 lessons of 50 minutes in blocks of two lessons). The themes included astronomy, dance, the newspaper, and statistics. The students mainly worked in pairs with Minitool 1 and 2. These experiments are described in Chapters 6 to 8.

3.7.3 Grade 8 experiment

The purpose of this experiment was to test newly developed instructional ideas and conjectures, in particular the idea of growing samples (Chapter 7) as a way to develop the notion of distribution in connection with that of sampling. This experiment was carried out in a *havo-vwo* school in the center of Utrecht, which had about 1,100 students, in October and November, 2001. The population was much more diverse than of the school where the seventh-grade experiments were carried out. The class was a *havo-vwo* class with 30 students (12 girls and 18 boys). The teacher had 11 years of teaching experience and also worked as researcher in mathematics education. We had weekly meetings about the planning of the next two lessons. The sequence consisted of ten lessons, as part of the mathematics lessons. Every odd-numbered lesson took place in a computer lab where the students worked in pairs with Minitools 1 and 2. Every even-numbered lesson was in a regular classroom without computer or a computer projector. Every week about four pairs of students, randomly chosen by the teacher, were interviewed for about 12 minutes to gain more insight

into their thinking about particular problems. Over the five-week period, all students had been interviewed except one who had been ill for most of the time. The day after the last lesson, semi-structured interviews were held with ten students for about 10 minutes per pair. The three assistants have also questioned all of the students about their interests, the minitools, contexts, and instructional settings. Some of the survey results are referred to in Chapter 10. Chapter 9 deals with this eighth-grade teaching experiment.

4 A historical phenomenology

4.1 Purpose

Various authors have suggested that studying the history of a topic is good preparation for teaching that topic (Dijksterhuis, 1990; Fauvel & Van Maanen, 2000; Freudenthal, 1983b; Gulikers & Blom, 2001; Radford, 2000). The obstacles that people in the past grappled with are interesting to teachers and designers because students often encounter similar obstacles. However, students also know things that people in the past did not know. What is the relationship between the historical development of statistical and mathematical concepts (phylogenesis) and the individual development of students (ontogenesis)? For a discussion of this question we refer to Radford (2000) for a mathematics education perspective and to Cole (1996) for a cultural psychology perspective. Here we confine ourselves to the question of what we can learn from a historical study for an instruction theory for early statistics education.
We used the RME heuristic of guided reinvention as a general guideline for the type of instruction we aimed for and we used a historical phenomenology as a method for studying the relation between the phenomena that were organized and the statistical concepts that historically arose for organizing such phenomena (2.1). Freudenthal (1983b) envisioned the following process of guided reinvention:

> The young learner recapitulates the learning process of mankind, though in a modified way. He repeats history not as it actually happened but as it would have happened if people in the past would have known something like what we do know now. It is a revised and improved version of the historical learning process that young learners recapitulate.
> 'Ought to recapitulate'–we should say. In fact we have not understood the past well enough to give them this chance to recapitulate it. (p. 1696)

Taking this cue from Freudenthal we decided to study the early history of statistics, in particular of averages, sampling, distribution, and graphs, which were the basic ingredients of the intended instructional sequence. In combination with a didactical phenomenology we could then conjecture on what "a revised and improved version of the historical learning process" for this particular topic might look like. As such this historical phenomenology also prepares the development of a hypothetical learning trajectory.
In this chapter,[10] we first discuss the method of our historical phenomenology, and then we study the average values (mean, midrange, mode), sampling, median, distribution, and finally graphs. The last section summarizes the insights that were most informative for the instructional design.

10. Sections 4.1 to 4.3 are based on Bakker (2003).

4.2 Method

When explaining what he meant by a phenomenology, Freudenthal (1983a) started his exposé as follows:

> I start with the antithesis–if it really is an antithesis–between *nooumenon* (thought object) and *phainomenon*. The mathematical objects are *nooumena*, but a piece of mathematics can be experienced as a *phainomenon*; numbers are *nooumena*, but working with numbers can be a *phainomenon*. (p. 28)

As Freudenthal wrote, there is a philosophical difficulty when distinguishing phenomena and concepts: it is hard to separate phenomena from concepts, since concepts also determine and influence how humans perceive the phenomena (e.g. Kant, 1787/1974). From a psychological perspective this distinction is also hard to make. For example, it is not exactly clear if and how humans' perception of colors depends on the terms that are available in their language (Anderson, 1995). For an educational purpose, it is still useful to try and separate phenomena and concepts, because students do not perceive the same phenomena as we do because of our understanding of certain concepts (an example of this is given in Chapter 9). Where statisticians see a clear pattern in a graph of a signal with noise (Konold & Pollatsek, 2002), students might just see a bunch of dots. Studying history can help us see certain phenomena through the eyes of people who did not have the same concepts and techniques as we have nowadays. By analyzing the historical process we expect to be able to identify different layers and aspects of concepts that seem to be fixed products nowadays. This attention to historical development may help us take a student's perspective and to better understand and guide the learning process.

The method of historical phenomenology requires identifying phenomena that have been organized by certain concepts and identifying concepts that have been applied to get a handle on certain phenomena. This historical phenomenology could be used to feed a didactical phenomenology, especially to find phenomena that challenge students to develop particular statistical methods or concepts. We have found no research literature in which this was already done for the statistical concepts that we were interested in, with the exception of Steinbring (1980), who writes about the development of chance and distribution with a didactical interest. Nor have we found historical phenomenologies that describe a systematic method.

Applying the method outlined above, we first collected as many historical phenomena with a statistical flavor as possible, mainly about center, sampling, distribution, and graphs (before 1900). Doing this we experienced several difficulties. One was that most histories of mathematics hardly pay any attention to the history of statistics. This need not be very surprising, because statistics arose largely from disciplines other than mathematics such as geodesy, astronomy, navigation, metallurgy, political arithmetic, medicine, anthropometry, biology, and social sciences. A second difficulty was that most historical studies of statistics start at around 1660,

which many authors mark as the start of probability and statistics (e.g. Hacking, 1975; Kendall, 1960), and focus on the nineteenth century (Porter, 1986; Stigler, 1986), whereas we had to go further back in time for the origins of the concepts we were interested in.

The next step, the selection of historical examples, was guided by the educational potential we saw in them; we only selected examples that could be regarded as preliminary stages of statistical notions with possible relevance for the design. For example, if an estimation of a number of years was reached by some method that could be interpreted as an intuitive version of an average, it was included in our phenomenology. A simple guess of a large number would not have been included. We also give an example from geometry to show that what might sound statistical ('arithmetic mean') need not be statistical. After many historical examples we formulate a hypothesis about students' learning, indicated with H#. Some of these hypotheses are revisited in the didactical phenomenology and in the retrospective analyses in the forthcoming chapters, but we have not been able to test them all.

Our methodology and purpose differ from what is common in historical science. In our historical study we do not describe the historical development of concept, but try to find sources of inspiration. We certainly do not want to suggest that 'the' historical development of statistical concepts was an accumulation of insights or a continuous refinement of concepts. As many authors nowadays note, revolutions and ruptures occurred in the history of mathematics and statistics as well (Gillies, 1992; Krüger et al., 1989). Nor do we want to suggest that students' learning always needs to follow the historical development. In this chapter we give a few examples in which students' cultural knowledge, for instance about surveys, makes it inefficient to follow the historical order. Although center, distribution, sampling, and graphs are highly interrelated topics, we address them separately for reasons of readability. There are six remaining sections: 4.3 on the average values (excluding the median), 4.4 on sampling, 4.5 on the median, 4.6 on distribution, 4.7 on graphs, and 4.8 is a summary of the most important results. In Chapter 5 we revisit several issues to investigate possible didactical consequences.

4.3 Average

If we use the term 'average' or 'average values' we refer to arithmetic mean, median, mode, midrange, and precursors of those measures of center. Because we study the early history of statistics as a source of inspiration for instruction to young students, we do not make technical distinctions between, for example, the mean of a population and the sample mean to estimate the center of a distribution. The terms 'center' and 'location' are also used informally for the center of a data set or a distribution.

4.3.1 Average values to estimate a total

The oldest historical examples that we considered relevant for the historical phenomenology all had to do with estimation of large numbers. Three of them are presented below to illustrate preliminary stages of several average values.

Example 1. Number of leaves on a branch

In an ancient Indian story, which was finally written down in the fourth century AD, the protagonist Rtuparna estimated the number of leaves and fruit on two great branches of a spreading tree (Hacking, 1975). He estimated the number on the basis of one single twig, which he multiplied by the estimated number of twigs on the branches. He estimated 2095, which after a night of counting turned out to be very close to the real number. Although it is uncertain how Rtuparna chose the twig, it could well be that he chose an average-sized twig, since that would lead to a proper estimation.

The educational potential we saw in this example was that such an implicit use of a representative value could be an intuitive predecessor of the arithmetic mean because one average number represents all other twig numbers and this average number is somehow 'in the middle' of the others. The choice is presumably made in such a way that what is counted too much on the one hand is counted too little on the other hand. This use of an average has to do, in our modern eyes, with compensation, balance, and representativeness. Even if Rtuparna did not use the method we think he used, the problem situation inspired our instructional design to let students reinvent such a method. (For a similar example of estimating a large number, the number of years between the first and last king of Egypt, see Rubin, 1968, or Bakker, 2003).

Rubin (1971) has found other old examples of statistical reasoning in the work of one of the first scientific historians, Thucydides (circa 460-400 BC). The following two quotations are from his *History of the Peloponnesian War*. The reader is invited to decide how he or she would translate these two excerpts into modern statistical terms.

Example 2. Height of a wall of Platea (Figure 4.1)

> (The problem was for the Athenians)... to force their way over the enemy's surrounding wall... Their method was as follows: they constructed ladders to reach the top of the enemy's wall, and they did this by calculating the height of the wall from the number of layers of bricks at a point which was facing in their direction and had not been plastered. The layers were counted by a lot of people at the same time, and though some were likely to get the figure wrong, the majority would get it right, especially as they counted the layers frequently and were not so far away from the wall that they could not see it well enough for their purpose. Thus, guessing what the thickness of a single brick was, they calculated how long their ladders would have to be...
> (Rubin, 1971, p. 53)

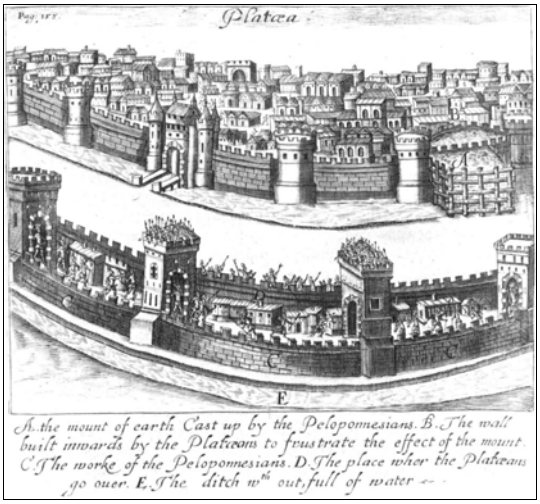


Figure 4.1: Walls of Platea as pictured in Hobbes' translation (1634) of Thucydides (1975, p. 189)

Example 3. Crew size on ships

> Homer gives the number of ships as 1,200 and says that the crew of each Boetian ship numbered 120, and the crews of Philoctetes were fifty men for each ship. By this, I imagine, he means to express the maximum and minimum of the various ships' companies... If, therefore, we reckon the number by taking an average of the biggest and smallest ships... (Rubin, 1971, p. 53)

We interpret example 2 as an implicit use of the mode, here indicated by "the majority," because "the majority" probably means "the most frequent value" and not necessarily "more than half." In this situation, the Greeks probably assumed that the most frequent number would be the correct one. To find the total height of this number of bricks, they supposedly needed another estimation: the expected or the average thickness of a single brick.

Example 3 also illustrates an estimation that is based on an average value. Thucydides possibly interpreted the given numbers as the extreme values, so that the total amount of men on the ships could be estimated by taking the average of these two extremes. In fact this is called the midrange, defined as the arithmetic mean of the two extremes. This technique of averaging the extreme values of the range to obtain the midrange can be justified if certain assumptions are defensible, for instance that the underlying distribution is approximately symmetrical.

Resuming, in these historical examples we encountered phenomena that were organized by predecessors of contemporary statistical concepts. In examples 1 and 2, a kind of average similar to the *arithmetic mean* was probably used. In example 2, we

can also recognize the *mode*. In example 3, Thucydides described a method that we can call taking the *midrange*. In these estimation examples these notions of average were not defined or used explicitly, although many mean values were known in those days (Heath, 1981). The median, however, was absent in the early examples. Eisenhart (1974), who investigated these issues in detail, has found no possible precursors to the median before 1599.
The conjecture that arises from these historical examples is the following.

H1. Estimation of large numbers could challenge students to use intuitive notions of average.

In Section 6.3 we describe how we tested and confirmed this conjecture in seventh-grade classes.

4.3.2 Mean values in Greek geometry

Explicit use of mean values and names for these values are found in ancient Greek mathematics. In Pythagoras' time, around 500 BC, three mean values were known, namely the harmonic, geometric, and arithmetic mean (Heath, 1981; Iamblichus, 1991). Only some 200 years later, at least eleven different mean values had been defined (Heath, 1981). For a historical phenomenology it is relevant to study the phenomena that gave rise to these concepts. It turns out that the theory of the three mentioned mean values was developed with reference to music theory, geometry, and arithmetic. We provide an example of the mean values in geometry to illustrate that these mean values were not used in a statistical way. Yet we can learn from the geometrical representation and the definitions of the mean values.
This example from geometry, a theorem of Pappus, illustrates that the Greeks studied the mean values for their geometrical beauty (see Figure 4.2) and not in a statistical sense. If in the semicircle ADC with center O one has DB $\perp$ AC and BF $\perp$ DO, then DO is the arithmetic mean, DB the geometric mean, and DF the harmonic mean of the magnitudes AB and BC (Boyer, 1991). This theorem does clearly not belong in a statistics course at the middle school level. Yet two aspects are important: the definitions of the arithmetic mean and the representation of magnitudes.

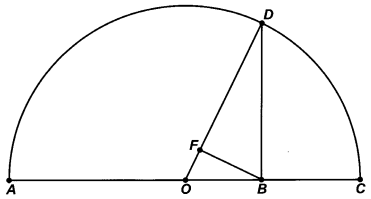


Figure 4.2: Theorem of Pappus on arithmetic, geometric, and harmonic mean.

In addition to the Greek definition of the arithmetic mean, Aristotle (384-322 BC) defined a philosophical form of the mean, the "mean relative to us." About the difference between the arithmetic mean and "the mean relative to us" he wrote:

> By the mean of a thing I denote a point equally distant from either extreme, which is one and the same for everybody; by the mean relative to us, that amount which is *neither too much nor too little*, and this is not one and the same for everybody. For example, let 10 be many and 2 few; then one takes the mean with respect to the thing if one takes 6; since 10-6 = 6-2, and this is the mean according to arithmetical proportion [progression]. But we cannot arrive by this method at the mean relative to us. Suppose that 10 lb. of food is a large ration for anybody and 2 lb. a small one: it does not follow that a trainer will prescribe 6 lb., for perhaps even this will be a large portion, or a small one, for the particular athlete who is to receive it; it is a small portion for Milo, but a large one for a man just beginning to go in for athletics. (*Nichomachean Ethics,* book II, chapter vi, 5; italics added)

The description "not too much and not too little" for the average is one that students used in all of the seventh-grade teaching experiments in the context of estimation (6.3).

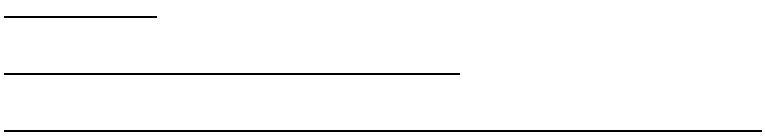


Figure 4.3: Greek representation of magnitudes as bars (2, 6, and 10)

In Greek mathematics, numbers and magnitudes were represented by lines. Aristotle's example with the mean of 10 and 2 represented in the Greek way (Figure 4.3) illustrates that Greek mathematics had a different form and aim than modern mathematics: it was highly geometrical and visual. This difference between Greek and modern mathematics can also be demonstrated by the difference in definitions of the arithmetic mean. The Greek definition, as we saw in the quotation of Aristotle, is as follows: the middle number b of a and c is called the arithmetic mean if and only if a-b = b-c. Note that this definition differs in formulation from the equivalent modern one, (a+c)/2, and that it refers to only two values. The Greek version shows that the mean is in between the two extremes and that it is difficult to generalize, whereas the modern version emphasizes the calculation and is easy to generalize. With the didactical phenomenology in mind, it is important to note that the Greek definition shows other qualitative aspects than the modern quantitative one. For example, we can immediately see from the Greek definition that the mean is halfway between the two other values. This feature is used in Greek astronomy for interpolation (Ptolemy, 1998), but we consider this application as non-statistical. Yet we highlight this 'in-

termediacy' aspect because students do not always realize that the mean is in between the extreme values (Strauss & Bichler, 1988). In a representation such as 4.3 students might be able to see that the part of the longest bar that 'sticks out' (compared with the middle bar) compensates the part of the shortest bar.

H2. The Greek bar representation might support the understanding that the mean is in between extreme values (intermediacy) and it might even scaffold a compensation strategy of visually estimating the mean.

In Section 6.7 we discuss how this conjecture was tested and confirmed in the teaching experiments.

4.3.3 Average, midrange, and generalization of the mean

In Section 4.3.1 we illustrated how the average was sometimes used implicitly in estimations and in Section 4.3.2 we conjectured that the Greek way of representing numbers by bars has educational potential for visual estimation of the mean. In the present section we describe how the average emerged from fair share in trade and insurance contexts, and that taking the mean of only two extreme values, the midrange, could be a predecessor of the arithmetic mean of more than two values in the context of science.

In the first millennium before Christ, the sea trade in the Mediterranean was lively (Plön & Kreutziger, 1965). During a storm, captains of small vessels with valuable merchandise sometimes needed to cut away the mast or throw some cargo overboard to avoid capsizing or to save the rest of the cargo. This act of throwing cargo overboard became known as the 'jettison' of cargo.

Een Tractaet
Van
AVARIEN,
Ghemaeckt by
QVINTYN WEYTSEN,
Raedt des Con. Maj^t. in den Hove van Hollandt.

AVARIE is ghemeyne contributie / van goeden inden Schepe bevonden / om te helpen draghen het verlies van eenighe Coop-lieden ofte Schippers goeden ghewillichlijck ghedaen / om lijf/schip/ende goet te saluveren.

Is te considereren / dat alle natien van herstenwaerts hoedanich die zijn / uyteren in't ghemeyne eenigh in haerlieden moederlijcker tale ende spraecke / dit woort Avarie: ende is 't selfde ghecomen uyt die van Griecken / alwaer groote frequentatie ende handelinghe van Coopmanschappe eertijdts plach te zijne à verbo βαρέω 'twelck beteeckent/onus ofte last/& adjectione privativa ἀ, quasi sine onere, unde ἄβαρις & inde communi lingua Avarie: soo seghtmen ghelijck te zijne/als een Schip ghewoppen hebbende arriveert sonder zijn volle last.

Figure 4.4: Part of the first page of a Dutch book on average by Weytsen (1641)

From about 700 BC, merchants and shippers agreed that damage to the cargo and the ship should be shared equally among themselves. What a merchant had to pay was called his 'contribution'. This idea became part of customary law and was written down in the 'lex Rhodia de iactu', the Rhodian law on jettison during the codification of Roman Law in 534. The basic principle in the Digest XIV.2.1 is as follows.

> The Rhodian law decrees that if in order to lighten the ship merchandise has been thrown overboard, that which has been given for all should be replaced by the contribution of all. (Lowndes & Rudolf, 1975, p. 3)

The rest of the text explains what should be done in specific situations and raises questions like, "In which proportion should compensation be paid?" Digest XIV.2.2.4 states that the equalized portion should take into account what the value of the saved and the lost cargo was. The number examples in the Latin texts are extremely simple and not very explicit. In Digest XIV.2.4.2 we read, for example:

> If therefore, for instance, two persons each had merchandise valued at 20,000 sesterces and one lost 10,000 due to water damage, the one with the saved merchandise should contribute according to his 20,000, but the other on the basis of the 10,000. (Spruit, 1996; translation from Latin and Dutch[11])

Old Dutch books on average (e.g. Weytsen, 1641) were not very explicit either (Figure 4.4). In search of more realistic examples of calculations we resorted to books of the nineteenth century that describe how to calculate averages (e.g. Arnould & Maclachlan, 1872; Hopkins, 1859; Van der Hoeven, 1854). These averages were calculated by a so-called 'average-adjuster', who was a kind of accountant. This must have been a serious profession, because there was even an 'Association of Average Adjusters' in England in the nineteenth and early twentieth century (Lowndes & Rudolf, 1975).

From this law-historical account we can track the development of the average in maritime law. Important for this historical phenomenology is that the average's origin is fair distribution and that proportions play a major role (cf. P11). But how did the term 'average' also come to signify the arithmetic mean?

The Oxford English Dictionary (Simpson & Weiner, 1989) writes that one of the meanings of 'average' in maritime law is "the equitable distribution of expense or loss, when of general incidence, among all the parties interested, in proportion to their several interests." In its transferred use it came to signify the arithmetic mean:

> The distribution of the aggregate inequalities (in quantity, quality, intensity, etc.) of a series of things among all the members of the series, so as to equalize them, and ascertain their common or mean quantity, etc. (...) the arithmetical mean so obtained.

From the examples in this section we see that this type of average originally arose

11. Translations in this thesis by AB unless indicated otherwise.

from the phenomena of fair share and insurance.

H3. Learning the mathematics that is involved in fair share and insurance (e.g. proportions and ratios) is good preparation for learning about the arithmetic mean. Fair share is also a suitable context to practice such skills (cf. Cortina et al., 1999).

Another possible precursor to the arithmetic mean is the *midrange*, which was used for example in Arabian astronomy of the ninth to eleventh century, but also in metallurgy and navigation (Eisenhart, 1974). Nowadays we model many observations and errors in those contexts with symmetrical distributions. Therefore, it is understandable that the midrange was used in those situations. Because the midrange was probably a precursor to the mean as a way to organize the center or estimate the true value, it might well be that students also use the midrange as a precursor to the mean.

H4. Students may use the midrange as a precursor to more advanced notions of average.

Not until the sixteenth century was it recognized that the arithmetic mean could be generalized to more than two cases: $\bar{a} = (a_1+a_2+...+a_n)/n$. Székely (1997) supposes that the invention of the decimal system by Stevin in 1585 facilitated such division calculations. This generalized mean proved useful for astronomers who wanted to know a real value, such as the position of a planet or the diameter of the moon. Using the mean of several measured values, scientists assumed that the errors added up to a relatively small number when compared to the total of all measured values. This method of taking the mean for reducing observation errors was mainly developed in astronomy, first by Tycho Brahe. From the late sixteenth century onwards, using the arithmetic mean to reduce errors gradually became a common method in other areas as well (Eisenhart, 1974; Plackett, 1970). This implies for our didactical phenomenology:

H5. Repeated measurement might be a useful instructional activity for developing understanding of the mean and distribution (cf. Konold & Pollatsek, 2002; Lehrer & Schauble, 2001). See also H12 and Section 10.4.

A question that arose was how we could benefit from the Greek definition and bar representation and still reach a general definition of the mean on *n* values (see Section 5.4).

4.3.4 The mean as an entity in itself

The historical examples until about the nineteenth century mostly had to do with approximating a real or best value, for example the number of leaves on a branch or

the diameter of the moon. In these old examples, the mean was used as a means to an end. It took a long time before the mean was used as a representative or substitute value as an entity in itself. The Belgian statistician Quetelet (1796-1874), famous as the inventor of *l'homme moyen,* the average man, was one of the first scientists to use the mean as the representative value for an aspect of a population. This transition from the real value to a representative value as a statistical construct was an important conceptual change (Porter, 1986; Stigler, 1986). What is relevant for the historical phenomenology is that there are several layers of understanding the mean as a representative value. We conjecture the following.

H6. Using an average value in estimations of large numbers and using the mean for reduction of errors are probably easier for students than understanding the mean as an entity in itself, that is as a representative value for an aspect of a population.

In Sections 5.1.6 and 5.1.7 we give some empirical support for this conjecture. It makes a difference if the mean as an entity in itself stands for a value that can exist or cannot exist. In 1877, Peirce–the same whom we revisit Chapters 8 and 9–wrote about this issue:

> In studies of numbers, the idea of continuity is so indispensable, that it is perpetually introduced even where there is no continuity in fact, as where we say that there are in the United States 10.7 inhabitants per square mile, or that in New York 14.72 persons live in the average house. [Footnote:] This mode of thought is so familiarly associated with all exact numerical consideration, that the phrase appropriate to it is imitated by shallow writers in order to produce the appearance of exactitude where none exists. Certain newspapers, which affect a learned tone, talk of "the average man," when they simply mean *most men*, and have no idea of striking an average. (CP 2.646)

4.4 Sampling

4.4.1 Estimation and sampling

Below the surface of the estimation examples, sampling issues also play a role. For instance, in the Indian story on estimating the number of leaves and fruit on a branch, the right twig had to be *chosen* to find an accurate average or representative value. Centuries later, John Graunt used a similar method of estimating the population of London, and Laplace to estimate the population of France (Bethlehem & De Ree, 1999), but they dealt more explicitly with sampling issues and reliability than we can infer from the examples in Section 4.3.1. Graunt knew that in parishes with reliable information about the population about three people died per eleven families per year. He also knew that there were about 13,000 funerals per year in London and he estimated the average family size as eight, which led him to 13,000 / 3 * 11* 8 is roughly 384,000 inhabitants of London. Also in Laplace's example, average values

were used to find a total number.

H7. More complicated estimation tasks than those of Section 4.3.1, such as those of Graunt and Laplace, might be useful to deal more explicitly with sampling issues.

Figure 4.5: Definition of a rod of 16 feet depicted by Köbel in 1535 (Mathematikunterricht 48(3), 2002).

Conversely, a total number can also be used to find a mean as a measure. Such an example of average in which sampling plays a role occurs in a geometry book by Köbel in 1535. Figure 4.5 shows how a rod of 16 feet should be determined by measuring the feet of sixteen men as they leave church (Stigler, 1999). This rod of 16 feet was to become a standard for other measurements in the community. In this simple measurement example we encounter several statistical issues. Were people in the past aware of the fact that the total of 16 feet is equal to 16 times the arithmetic mean of the lengths of these 16 feet? Probably not because the arithmetic mean was only defined later for more than two values. How was the sample taken? We may assume that there was no size-based criterion for selection. Although many scientists in 1535 still assumed that combining observations would amplify the errors instead of reduce them (Stigler, 1986), this example seems to be an intuitively clear way of combining measurements to reduce variation. Did the inventors of this method realize that they benefitted from compensation of errors? For the purpose of our historical phenomenology it is not necessary to know this; we simply use the example as a source of inspiration for instructional activities (6.3, 6.4, and 6.9). As the examples show, variation and sampling issues often underlie seemingly simple problems concerning averages.

H8. Examples similar to those in this section may be used to let students think about the arithmetic mean in close connection to variation, measurement, and sampling.

4.4.2 Decision-making

There are examples of sampling that reveal yet another origin of statistics: decision-making. We give a few examples from Jewish Law. Jewish Law deals mainly with

social, ethical, and ritual duties and is considered a rational pursuit: although rabbis accepted divine guidance, they insisted on rational methods in coming to decisions. Rabbis had to decide, among other things, how inheritances had to be distributed and whether food was kosher. If, for example, 9 out of 10 shops in a city sold kosher meat and someone found a piece of meat in that city, a rabbi could advise to consider it kosher (Rabinovitch, 1973). We interpret this as follows. If from a sample of 10 shops 9 sell kosher food, this proportion gives an indication of the chance that an arbitrary piece of meat in the city is kosher. Thus proportional or multiplicative reasoning plays an important role in the relation of sample to population.
Another decision-making example concerning multiplicative reasoning and sampling concerns the question of whether an epidemic has taken place, which is relevant to know (as it is today) for undertaking particular steps.

> A town bringing forth five hundred foot-soldiers like Kfar Amiqo, and three died there in three consecutive days - it is a plague... A town bringing forth one thousand five hundred foot-soldiers like Kfar Akko, and nine died there in three consecutive days - it is a plague; in one day or in four days - it is not a plague. (Rabinovitch, 1973, p. 86)

In this example three points are interesting for the historical phenomenology. First, we see that rabbis reasoned proportionally to the total population. Second, a kind of sampling was used: the amount of foot soldiers was used as an indicator of the total population, probably because foot soldiers formed a constant percentage of the population. Third, the rabbis seemed to know typical or average death rates and they took into account how the deaths were distributed over the consecutive days.

H9. When making data-based decisions, multiplicative reasoning is an essential skill in dealing with samples versus populations (cf. P11 in Section 2.3).

4.4.3 Quality control

Apart from estimation and decision-making there are several other origins of statistics. One of them is quality control. In this section, we discuss an old secular example of sampling and quality control: the trial of the Pyx (Stigler, 1977). This trial took place at the Royal Mint of Great Britain where gold and silver coins were made. Starting from the twelfth century, every day one of the coins was put in the Pyx, which was a box in Westminster Abbey. After a few months or years, the Pyx was opened and the coins were investigated on weight and pureness. Not the single coins but the whole box was weighed and a sample of coins was melted to investigate the pureness of the coins. If the coins turned out to be good, this fact was celebrated with a banquet; otherwise the coin makers were punished. This is the first clear example of quality control we have found that is based on sampling inspection.
In Section 2.2 we wrote that we were in search for coherent knowledge of the key concepts of statistics. What we can learn from this historical phenomenology for the didactical phenomenology is that the use of averages often involves sampling issues

(4.4.1). This close link to sampling indicates one of the difficulties of learning the mean, but the link can also be used to teach sampling issues from what students already know about averages. Another thing we conjectured from this example is that sampling as carried out here, one coin per day, might be an intuitively clear way of sampling to students as well (but see Section 6.9).

H10. Problem situations similar to the trial of the Pyx may challenge students to reinvent simple sampling methods. Randomness is implicit in the trial of the Pyx.

H11. Moreover, such problems may be used to reinforce a meaningful relation between average and total, which in turn can form the basis for insight into the relation of sample and population.

4.4.4 Random sampling

From the examples in the previous sections on sampling we can infer that sampling can be used for different purposes such as finding a total number, finding a measure based on a total, and making a decision. Historically the next stage was to use sampling for getting information about a population. There are different reasons to use sampling. An important motive for the Central Bureau of Statistics in the Netherlands was to reduce the costs of its studies (Bethlehem & De Ree, 1999), and often it is also impossible to measure the whole population. Yet is sampling a relatively recent accomplishment.

Censuses were held both in ancient China and Egypt. Famous, of course, is the Roman census of Caesar August that is known from the biblical story about Jesus's birth. Incas (1000-1500) recorded information about their people, homes, llamas, marriages, and young men that could be recruited for the army. Until late in the nineteenth century, only integral surveys were carried out because other methods were considered unreliable and discriminatory. It was considered unfair to take observations of certain human beings into account and replace those of others by calculations. This last feeling of resistance is understandable if we make a comparison with voting. Today most of us would also protest if we were not allowed to vote for a new government and a sampling method were to be used instead. Yet statisticians argue that a good sampling method is more reliable than a self-selection which results from a non-obligatory call to vote (De Mast, 2002).

A new period for statistics started in 1895 when the Norwegian Kiaer presented his 'representative method', which implied the deliberate selection of a representative sample, for instance as many men and women, from cities and villages, of all ages, and so on (stratified sample). It was not until 1903, however, that the International Statistical Institute accepted this method provided that the selection was carefully described. In 1906, Bowley proposed to use a process of drawing lots. The two methods coexisted until Neyman (1934) was finally able to prove that random sampling

was superior to Kiaer's method.
One of the underlying conceptual difficulties of sampling is its close link to probability as we already hinted at in the section on decision-making. In *The Probabilistic Revolution* (Krüger et al., 1989) different authors underline the conceptual shift from determinism to indeterminism that was made at the end of the nineteenth century, and which proved crucial to the development of statistics and probability theory (see also Hacking, 1990; Porter, 1986; Stigler, 1986).
In contrast to people in and before the nineteenth century, students today are acquainted with surveys, which are now culturally accepted, and students might also know about random numbers from computer games. This means that students need not exactly follow the historical development of sampling, but it could still be that students think that everybody should be measured in some cases. From this historical outline we can distinguish different levels of understanding sampling.

H12. If the unit of thought or focus of attention is a concrete object such as a coin and if there is little variation, students may reinvent sampling methods. However, if the unit of thought or object of interest is a whole population that is influenced by multiple variables, students probably prefer stratified sampling to random sampling, because it gives the suggestion of having control of the sample.

Before the seventh-grade teaching experiments we had mainly paid attention to the average values and to sampling. The historical study of distribution for example had not yielded very much. The teaching experiments in grade 7, however, urged us to reconsider the history of the median, distribution, and graphs. Because we preferred to keep the historical phenomenology of the different concepts in one chapter, the history of these concepts and graphs is discussed in the next sections. As a consequence, the hypotheses in those sections were formulated only after the seventh-grade teaching experiments and they did not play an explicit role in the hypothetical learning trajectory of these experiments.

4.5 Median

After the seventh-grade teaching experiments we had two reasons to investigate the history of the median more carefully. First, the seventh-grade students had more problems with the median as a representative value than we had expected, even after taking the results of the Nashville team into account (R16).[12] To understand students' problems with the median we carried out a conceptual analysis of it. Second,

12. Cobb (personal communication, February 18, 2003) and Gravemeijer assume that students' problems with the median were due to the design (see also Cobb, McClain, & Gravemeijer, 2003), but we conjecture that there are conceptual problems with the median that are not easy to overcome.

during instructional design we could not find phenomena that beg to be organized by the median as a measure of center or a representative value. So far, our historical study had not helped us much further because we had hardly found any examples of the median before about 1840. Why did the median arise so late? Is it because it is a difficult concept? What was it used for? Can we find clues for instructional design? The historical phenomenology we undertake in this section is thus meant to find out why the median is difficult for students and to find phenomena that require using the median. Unfortunately, there turned out to be very few historical studies of the median; its history is mostly buried under bigger issues such as the normal distribution, Bayes' theorem, the central limit theorem, and the method of least squares. Monjardet (1991) describes the median's history with an interest in metric spaces; Godard and Crépel (1999) concentrate on the median's statistical characteristics after 1750; and Harter's *Chronological annotated bibliography on order statistics* (1977) provides a list of articles on order statistics, some of which deal with the question of whether the mean or median is a better measure of center (this was around 1900). Hence we had to do our own historical study of the median. We used the same method as for the previous sections.

We organize this section by focusing on the phenomena and contexts in which the median arose. It turned out that the median mostly emerged as a differentiation of the mean, and it was often the alternative measure that appeared less viable. In metaphorical terms, Ms Median appeared to be the stepsister of Ms Mean.

4.5.1 Theory of error

The most important context in which the average values were used was the theory of error (Sheynin, 1996). The problem at first was to find the assumed true value from the available observations, later to find the best estimate of such a value. The Greeks often used a value that fitted the theory instead of real observations to "save the phenomena," as they called it (Pannekoek, 1961; Steinbring, 1980). Another method was to choose a value that seemed reliable, for example from a middle cluster or from values measured under favorable conditions (cf. Section 7.2). As we mentioned before (4.3.3), the midrange was used as well, for instance by Arab scientists in the ninth to eleventh century. It was not until the late sixteenth century, however, that the mean of more than two values was defined (4.3.3). Tycho Brahe seems to be the first to use the mean for reducing error and combining observations (Placket, 1970).

The first possible instance of the *median* that Eisenhart (1974) has found was in a book by Edward Wright of 1599 on navigation. Wright wrote about the determination of location with a compass (note that the letters 'u' and 'v' were used differently in those days):

> Exact trueth amongst the vnconstant waues of the sea is to bee looked for, though good instruments bee neuer so well applied. Yet with heedfull diligence we come so neare the trueth as the nature of the sea, our sight and instruments will suffer vs. Neither if there be disagreement betwixt obseruations, are they all by & by to be reiected; but as when many arrows are shot at a marke, and the marke afterwards away, hee may bee thought to worke according to reason, who to find out the place where the marke stood, shall seeke out the middle place amongst all the arrowes: so amongst many different obseruations, the middlemost is likest to come nearest the truth.
> (Eisenhart, 1974, p. 52)

It is not certain that Wright really meant the median, since he gave no numerical examples. Eisenhart argued that since Wright wrote "Neither... are they all by & by to be reiected" it is possible that he recommended the middle-most observation, the median, and not the middle place, the midrange, since then most observations would not be used. Even if this is a real example of the median, it is just a solitary example, and certainly not an indication of a common practice of using the median in navigation or any other context.
A clearer example of the median, in the context of measurement errors, is found in the work of Boscovich (around 1755). The interesting point of his work for the history of the median was the set of conditions he proposed in the search for true values, in particular a line of best fit through observations. One of these conditions was that the sum of absolute errors should be minimal; in our notation: $\Sigma|x_i\text{-}x|$ is minimal. This condition turns out to be equivalent to the statistical median (David, 1998b; Eisenhart, 1977), which can be proven with differentiation. Note that the condition that the errors should add up to zero is equivalent with the arithmetic mean: $\Sigma(x_i - a) = 0 \Leftrightarrow \Sigma x_i / n = a$.

H13. If the theory of errors (e.g. repeated measurements) is taken as a context for developing statistical ideas of center and distribution, it may be advantageous to let students formulate their own intuitions about the distribution of errors. Do the errors add up to zero? Is the chance that the measurement is too small equal to the chance that they are too large? Do errors occur symmetrically?

In fact, discussions on the mean in error theory led to the development of the concept of distribution (4.6).

4.5.2 Probability

Another context in which the median arose as a counterpart of the mean was probability theory. We give three examples of how the median is connected to probability. The first is a paradigmatic example of how an intuition of something, in this case a middle value, became differentiated into two concepts, namely median and mean life time. The second example is about birth rates and deals with quartiles and the interquartile range. The third example stems from Legendre and Laplace, who dis-

tinguished two possibilities of finding a true value, one of which we would now call the median.

1. In 1669, the Dutch brothers Chistiaan and Lodewijk Huygens had an informal correspondence about their father's life expectancy and about life expectancy in general. In 1662, just a few years earlier, they had received the famous *Bills of Mortality* by John Graunt and with the tabular data in the book they calculated their father's and their own chances, which then evoked a flow of new mathematical problems (Véron & Rohrbasser, 2000). The brothers continued to write each other on annuities and life insurance based on these mortality tables, but they disagreed about certain calculations. It was Christiaan who realized that there was a difference between expected remaining life time and the life time that half of the people would reach. On November 28, 1669, he wrote to Lodewijk:

> There are thus two different concepts: the expectation or the value of the future age of a person, and the age at which he has an equal chance to survive or not. The first is for the calculation of life annuities, and the other for wagering. (C. Huygens, 1895, Volume 6, letter to Lodewijk Huygens; translation from French: Hald, 1990, p. 106)

Christiaan made a graph from which we can read the median life time; this graph was one of the first line graphs ever (Tufte, 2000, 2001; see also Section 4.7). In Figure 4.6 we can see that a 20-year-old person (A) had a median life time of 36 years: take the half of AB and find CD further in the graph. The French terms that Christiaan used for what we now call median life time were *apparence* (likeliness) and *vie probable* (probable life), since the person has equal chance to survive to this age or not. The chance of a half appears a natural point to look at, though it was not very useful except for chance-like problems such as wagering. More useful was what Christiaan called *espérance* and what we now call mean or expected life time. This is also what Johan de Witt and Jan Hudde used for the life annuity calculations two years later.
We conclude for the historical phenomenology that the phenomenon of predictions about life times asked for a distinction between mean and median life time due to skewed distribution (Stamhuis, 1996). The mean life time was useful for annuities, but the median was only useful for wagering.

H14. In this context of life times, that is in a skewed distribution, the median and mean refer to different intuitions of center. The median is connected to probability theory, in particular to halves, and the mean to expectation. When aiming at the median, it is worth trying to design problem situations in which it is reasonable to look at halfs or compare halves, for instance in chance situations.

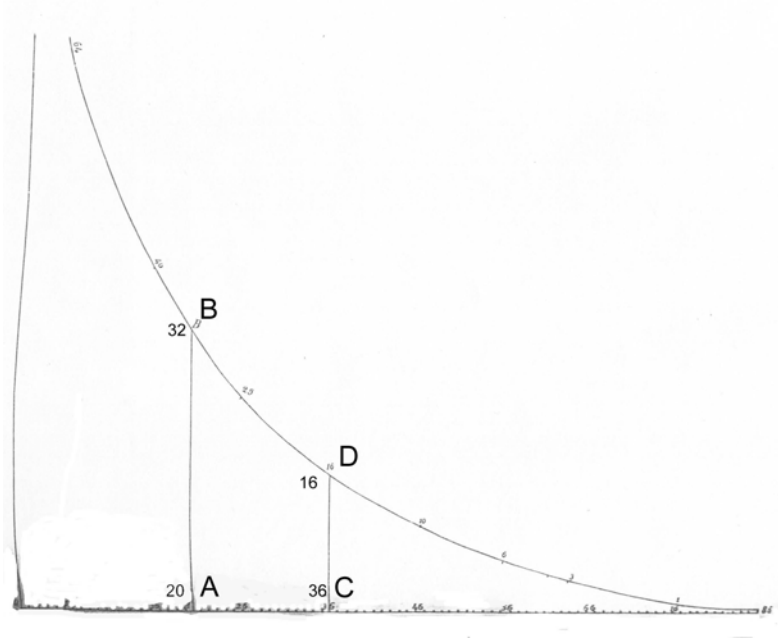


Figure 4.6: Huygens' theoretical line graph of mortality data (Huygens, 1895, between page 530 and 531). The letters and numbers in the original are enlarged for readability.

2. The second context in which we see a connection between probability and the median is birth rates. It brings us to quartiles and the interquartile range.
Mathematicians such as De Moivre, Stirling, and Daniel Bernoulli studied birth phenomena with binomial distributions, not with real data. Bernoulli, for example, wanted to calculate the probability that in a binomial distribution the variable appeared between two limit values. He assumed that somewhere 2N children were born, with equal chances for boys and girls. The essential point for our median story is that he then raised the question of what the limit values were that would delimit half of the cases. A hundred more boys could be indicated by +100, 24 more girls by –24. With 2N=20,000 he found that this limit value was 47 ¼ at either side of N. In general, he wrote, it is 0.4725√N, which is close to the value 0.4769√N that is derived from the normal distribution (Hald, 1990). For the middle range between the limit values Bernoulli used the term *status medius*; and for the middle limits the term *limites medii* (Bernoulli, 1982, pp. 220, 385). His *status medius* is the same as the interquartile range and the *limites medii* are the same as the first and third quartiles. With a modern view we might interpret the middle position as the median here–

equal numbers left and right–and not the mean.
What is striking in this context of dealing with birth rates is that the median and the quartiles are more apparent than the mean and modulus (this is a precursor to the standard deviation[13]; see Walker, 1931). The median is seen as the exact balance of boys and girls. Bernoulli (and later Galton) would probably have been surprised if they had heard that nowadays students learn the 68.26% rule for standard deviations in normal distributions, and do not work with the interquartile range of 50%.

H15. Quartiles and the interquartile range are intuitively clearer measures of variation than standard deviations (cf. P8).

3. The third example of the connection between the median and probability is the following. Legendre, Laplace (1812/1891), and their contemporaries used the term *milieu de probabilité*, the middle of the probability, which is a suggestive name for the median in the context of probability functions.
Cournot (1843) was the first to use the term 'median' (*valeur médiane*) for this value (Bru, 1984; David, 1995, 1998b; Stigler, 1986). He defined the median as the value x_0 for which the distribution function F was $F(x_0)=½$ and he explained that it is the value for which the area under the graph is the same on the left and on the right (Figure 4.7). Furthermore, he wrote:

> Two players, one betting of the value smaller than x and the other larger than x, would bet with the same chances. With a very big number, the quotient of the larger (or smaller) values than x and the total number of values will not differ much from the fraction ½. (Cournot, 1843, p. 83; translation from French)

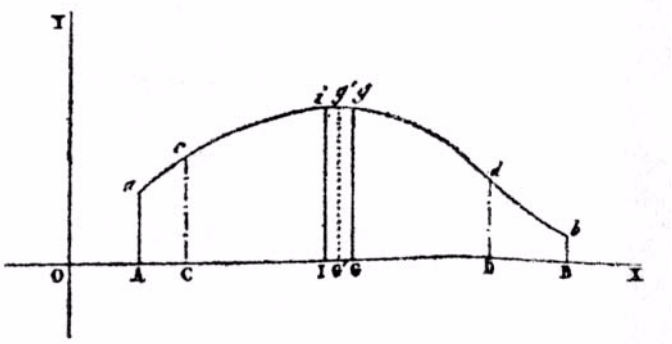


Figure 4.7: Graph from Cournot (1843, Figure 17) with the median in a skewed distribution.

13. The name 'standard deviation' was introduced by Karl Pearson at the end of the nineteenth century (David, 1995; Walker, 1931).

H16. One way to visually estimate the median in a dot plot such as in Minitool 2 is to look for which value the areas on the left and right are the same (see Section 10.5).

The central point of this section is that the median and quartiles are closely related to probability theory, especially with the chance of a half.

4.5.3 Ease of calculation and ordinal data

In 1874 Gustav Theodor Fechner (1801-1887) used the median, the *Centralwerth,* in an attempt to describe many sociological and psychological phenomena with methods that had proven to be useful in astronomy. He advocated the ease of calculation of the median, but he also had more theoretical reasons for using other measures of center than the mean, which we address in Section 4.5.5.
Francis Galton used the English term 'median' for the first time in 1882 (David, 1995) and caused the breakthrough of the concept (Godard & Crépel, 1999). As happens often in the history of mathematics and statistics (Bissell, 1996; Dijksterhuis, 1950), Galton knew the concept before he used this particular term. Before 1882 he used other terms including the *middle-most* value (1869) and the *medium* (1880), and in a lecture in 1874 he gave the following description:

> The object then found to occupy the middle position of the series must possess the quality in such a degree that the number of objects in the series that have more of it is equal to that of those that have less of it. (Walker, 1931, p. 87)

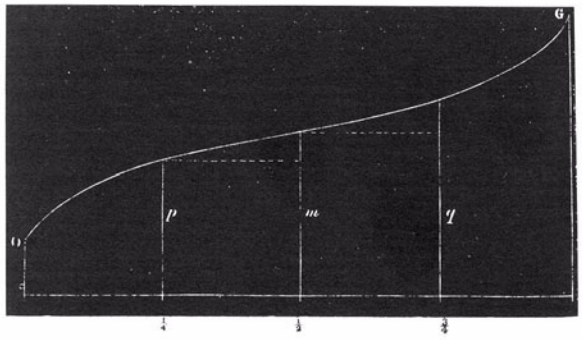


Figure 4.8: Galton's graph of the normal distribution with quartiles p and q, and median m (Galton, 1875, p. 36)

In a graph from 1875 he indicated the median and quartiles with the letters *p*, *m*, and *q* and the fractions $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$, but not with names (Figure 4.8). We come back to this graph in Section 4.6 on distribution.

H17. A five-number summary of extreme values, quartiles, and median may be a suitable way for students to characterize distributions, once students know the median and quartiles as measures of center and spread.

Important reasons for Galton to use the median were its ease of calculation and its intuitive clarity (Stigler, 1973). Most phenomena Galton studied were roughly symmetrical, so the median would not differ much from the mean, which is laborious to calculate. Throughout his book *Natural Inheritance* (1889) he therefore used the median M and quartile distance Q, with $Q = 1/2\ (Q_3 - Q_1)$, Q_1 being the first and Q_3 the third quartile (in the modern terminology). He rarely mentioned the mean and modulus ($\sqrt{2}$ times the standard deviation), probably to reduce calculation efforts and not scare away scientists without the necessary statistical background.
Apart from ease of calculation and intuitive insight, yet another reason for using the median could have played a role. Galton studied variables that he measured in an *ordinal* way. And indeed, with ordinal data the mean cannot be calculated, in contrast to the median.
What struck us in the historical study was that ordinal data are so rare, apart from the paradox of Borda and other theoretical voting problems that we interpret as non-statistical (Condorcet, 1785; Crépel & Godard, 1999; Goddijn, 1988; King, 1963; Monjardet, 1991). Galton seems to be one of the first to study real ordinal data, for instance in the context of intelligence and reputation. From the historical overview we conjecture the following.

H18. Although ordinality is a statistical reason to use the median as a measure of center, contexts with ordinal data are not very suitable to help students understand the median.

4.5.4 Robustness

Over the last centuries scientists have been concerned with the sensitivity of the mean to outliers, and have proposed different procedures that were more 'robust' as Box called it in 1953: trimmed means, weighted means, averaging different average values, but also the median (Stigler, 1973, 1980). Francis Ysidro Edgeworth (1845-1926), a younger contemporary of Galton, preferred the median to the mean because of its insensitivity to outliers, probably due to his interest in economics, which has less regular data than is common in astronomy for instance, and he was not the only one to prefer the median (Harter, 1977). Nowadays, the median's resistance to outliers is one of the major reasons to use it, especially when the data are irregular as is common in social sciences and economics.

H19. Using irregular data with outliers can motivate students to reason with the median instead of the mean as a measure of center, provided students already know about outliers and measures of center.

4.5.5 Skewed distributions

For a long time, distributions of error were assumed to be symmetrical. In 1838, Bessel was probably the first to doubt the assumption of symmetry (ESS, 1998; Steinbring, 1980). In contrast to most other scientists of his time, Fechner (1874) even assumed that *most* distributions of data were asymmetric. It turned out that the median minimizes the sum of absolute deviations to the first power and the mean the sum of deviations to the second power. Consequently, both measures of center are special cases of Fechner's so-called *Potenz-mittelwerthen*, the values that minimize the sum of the deviations to the n-th power: $\Sigma|x_i\text{-}x|^n$ is minimal. Fechner used these generalized measures of center to describe the skewness of distributions. Edgeworth also used the difference of mean and median, divided by a normalizing factor, and this measure is still used today as an indication of skewness (ESS, 1998; Stigler, 1986).

As Tukey (1977) pointed out decades later, the median is also useful in the five-number summary of unimodal distributions, consisting of the minimum value, first quartile, median, third quartile, and maximum value. In fact, this five-number summary is the basis of the box plot. The median is especially useful as a measure of center in asymmetric distributions, because it is far less influenced by extreme values than the mean (the median is more 'robust').

For the Nashville team, a reason to use the median was that it tends to be closer to the majority of a unimodal data set than the mean (R16). Another reason was that the median plus quartiles seemed easier than the mean plus standard deviation (P8). And third, the median and quartiles seem more appropriate when describing skewed distributions than mean and standard deviation (4.5.2).

H20. Skewed distributions can be used to show the usefulness of the median.

4.5.6 Summary of the median's history

The mean was used for reducing error from the late sixteenth century onwards, but the median was developed relatively late. We summarize a few examples. Expected life time was a useful element in calculating life annuities, whereas the median life time was considered to be just "for wagering," as Christiaan Huygens wrote. For Daniel Bernoulli the quartiles (*limites medii*) were evident values to look at, but the standard deviations won in the nineteenth century.[14] In the theory of error the mean

14. There are of course exceptions: in 1910, the Dutch botanist Tine Tammes preferred median and quartiles (De Knecht-van Eekelen & Stamhuis, 1992).

became a popular measure for combining observations and reducing error, but the median was not used until it was acknowledged that distributions can be skewed. The contexts in which such distributions became apparent were for example the social sciences and economics (Crépel & Godard, 1999). Phenomena in these fields are far less regular than in astronomy and physics, so a measure of location that is less sensitive to outliers, such as the median, is useful. Despite the efforts of scientists like Cournot, Fechner, Galton, and Edgeworth, the median was neglected and the mean favored.

Nowadays, the median is used in order statistics, since the mean cannot be used for ordinal data. The median is also used in robust statistics, since it is far more robust than the mean. Since statistics is applied in more and more areas with irregular data, the median has become more popular (Portnoy & Koenker, 1997). The question remains, however, whether Ms Median as the stepsister of Ms Mean will ever turn out to be Cinderella.

4.6 Distribution

For reasons of readability, we organized the historical phenomenology of average values, sampling, and distribution into different parts, but as we argued in Section 2.2 and as the examples in previous sections show, all these concepts are intimately interwoven. Estimation using average values has to do with sampling and the median examples often involve distribution issues.

In the eighteenth century, the concept of distribution arose from the theory of errors, when the arithmetic mean as a method to reduce errors was still a topic of debate. Due to the impossibility of determining individual errors, one had to look at the relation between the errors.

> Measurements, and functions of measurements, such as their arithmetic mean, are not amenable to mathematical theory, (...) as long as individual measurements are regarded as unique entities, that is, as fixed numbers y_1, y_2, … A mathematical theory of measurements, and of functions of measurements, is possible only when particular measurements y_1, y_2,... are regarded as instances of hypothetical measurements Y_1, Y_2,... that might have been, or might be, yielded by the same measurement process under the same circumstances. (ESS, 1998, p. 531)

In 1756 Simpson made this shift to looking at the relation between errors when he used simple probability functions to argue that the mean of several observations was better than a single observation. The first distribution of errors he proposed was a discrete uniform distribution, that is with equal probabilities for all values $-v$, $-v+1$,...-1, 0, 1,..., v. Next, he assumed a discrete isosceles triangle distribution with probabilities proportional to 1, 2,..., $v-1$, v, $v+1$, v,..., 2, 1, from which he obtained a continuous isosceles triangle distribution one year later (see Figure 4.9).

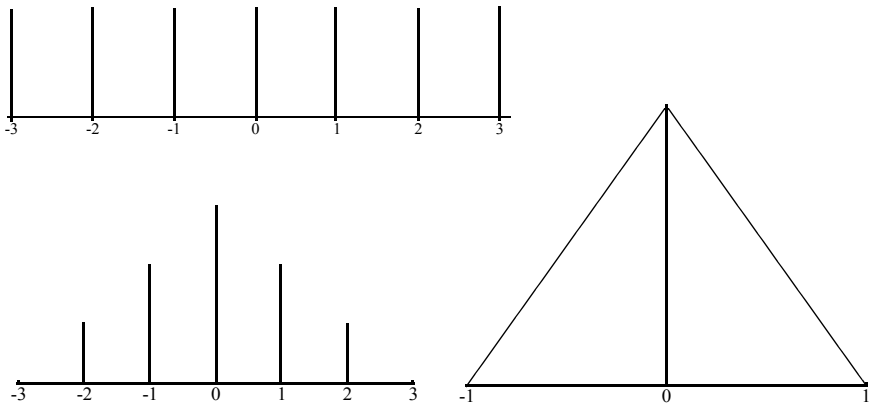
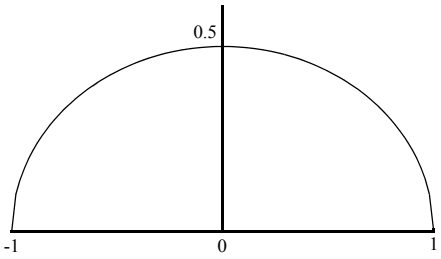


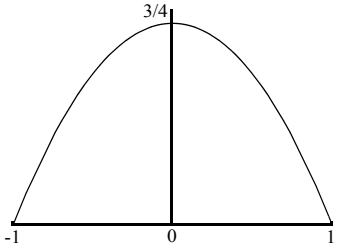
Figure 4.9: Distributions proposed by Simpson: discrete uniform (1756), discrete isosceles triangle (1756), continuous isosceles triangle (1757) (after ESS, 1998)

Quickly after Simpson had launched his idea of probability distributions, other scientists proposed alternative laws of error. Among them were Lagrange, Lambert, Daniel Bernoulli, Laplace, and Gauss. Note that the analytic expressions in Figure 4.10 are a modern accomplishment. Lambert, for instance, introduced his method of maximum likelihood without ever expressing his error-frequency distribution in a functional form (ESS, 1998).



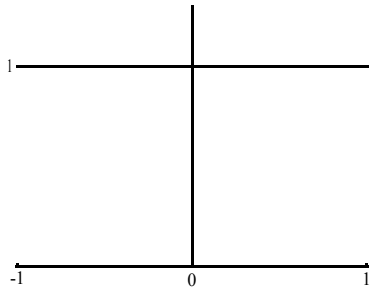
Lambert (1765): flattened semicircle

$$f(x) = \frac{1}{2}\sqrt{(1-x^2)} \qquad -1 < x < 1$$

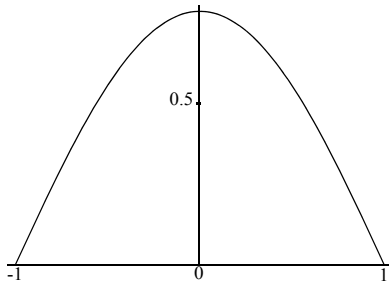


Lagrange (1776): continuous parabolic

$$f(x) = \frac{3}{4}(1-x^2) \qquad -1 \le x \le 1$$

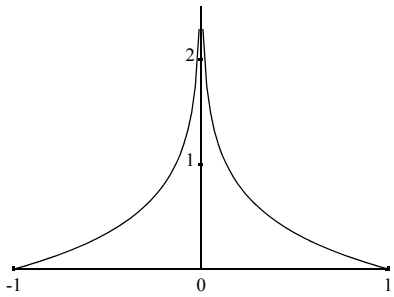


Lagrange (1776): continuous uniform



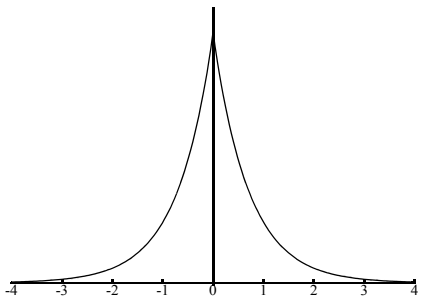
Lagrange (1781): cosine function

$$f(x) = \frac{\pi}{4}\cos\left(\frac{\pi x}{2}\right) \qquad -1 \le x \le 1$$



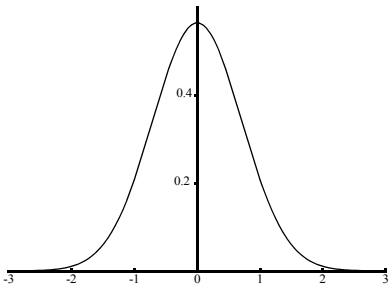
Laplace (1781): log function

$$f(x) = \frac{1}{2}\log\frac{1}{|x|} \qquad -1 \le x \le 1$$



Laplace (1774): double exponential

$$f(x) = \frac{1}{\sqrt{2}}e^{-\sqrt{2}|x|}$$



Gauss (1809) and Laplace (1810): law of error or normal distribution

$$f(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$$

Figure 4.10: Chronological overview of distributions of error that were proposed from Simpson to Gauss and Laplace (adapted from ESS, 1998)

H21. Students can reason about the shape of distributions without being bothered by analytic expressions (R14). They may implicitly assume distributions to be symmetrical.

Mathematically, there are many ways in which the normal distribution can arise. Historically the first way, and still a very common one, is to obtain the normal distribution as the limit of the binomial distribution *bin(n, p)* with *n* to ∞. This is a result of the De Moivre-Laplace limit theorem, which is a special case of the central limit theorem transpiring that in many cases the sum of a large number of independent random variables is approximately normally distributed.
In the context of people's height, we can think of many factors that influence this including their parents' height, their diet in their youth, their age, and their sports history. Even if these factors themselves are not normally distributed, their sum roughly is. This explains why so many phenomena can be described by the normal distribution (Sittig & Freudenthal, 1951; Wilensky, 1997).
This distribution and its curve are known under many names (Stigler, 1999) including 'the law of error', the 'frequency law', the 'Gaussian curve', 'Laplace-Gauss' (mainly in the French literature). One name Stigler does not mention is the 'De Moivre distribution', which Freudenthal (1966b) used because De Moivre was the first to define this function. An immediate result of Gauss's work with the normal distribution was that astronomers were able to find the planetoid Ceres in the sky again (Steinbring, 1980). Quetelet then used methods that had proven successful in astronomy for anthropometrical purposes, and modeled phenomena such as the chest sizes of Scottish soldiers with a binomial distribution with the curve of the normal distribution superimposed as the so-called 'curve of possibility'. In his tracks, Galton used the normal distribution ('normal scheme') for his studies in human faculties and inheritance. A famous quote is:

> It is difficult to understand why statisticians commonly limit their inquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect from Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once. An Average is but a solitary fact, whereas if a single other fact be added to it, an entire Normal Scheme, which nearly corresponds to the observed ones, starts potentially into existence. (Galton, 1889b, p. 62)

Galton realized that he needed just two numbers for describing the whole distribution: the median and the quartile distance.

> If we know the value of M [median or mean] as well as that of Q we know the entire Scheme [normal distribution]. M expresses the mean value of all the objects contained in the group, and Q defines their variability. (Galton, 1889b, p. 61)

Being interested in tribes–he traveled in Africa in 1851 (Galton, 1889a)–he recommended anthropologists to ask chiefs to arrange their people in order of height. Determining the first quartile, median, and third quartile would be sufficient statistics to describe the whole height distribution of the tribe (Walker, 1931, p. 84). Compared to the earlier large tables with observations, this use of a distribution was indeed a major leap forwards (Figure 4.11). In fact, this is why students need to develop such notions in relation to distributions:

> Over-minuteness is mischievous, because it overwhelms the mind with more details than can be compressed into a single view. (1889b, p. 36)

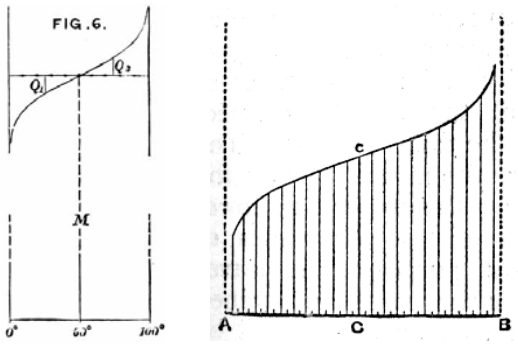


Figure 4.11: a) Another ogive-shaped graph with quartiles by Galton (1889b, p. 40); b) one with 21 hypothetical data values (1883, p.51)

Galton was really impressed by the normal distribution (‘law of frequency of error’), which inspired him to write poetic phrases:

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along. (Galton, 1889b, p. 66)

Several scientists compared Gauss's law of error with observed frequencies in mass phenomena and considered the agreement as good. Among these scientists was C.S. Peirce (NEM III), who collected many data values on a young man's reaction time on consecutive days. About the curves of Figure 4.12 he wrote:

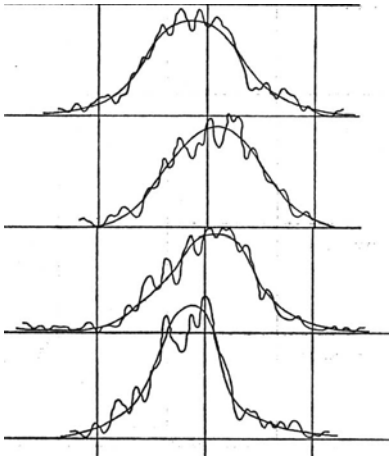


Figure 4.12: One page of Peirce's curves that, according to Peirce, differed little from the normal distribution: reaction time of an individual on consecutive days (Peirce, NEM III, p. 676)

> The curve has, however, not been plotted directly from the observations, but after they have been smoothed off by the addition of adjacent numbers in the table eight times over, so as to diminish the irregularities of the curve. The smoother curve on the figures is a mean curve for every day drawn by eye so as to eliminate the irregularities entirely. It was found that after the first two or three days the curve differed very little from that derived from the theory of least squares [the normal distribution].
> (NEM III, p. 659)

The belief in the general applicability of the normal law was ubiquitous and slow to die. Lippmann, a French physicist, said to Poincaré (1854-1912) about this:

> All the world believes it firmly, because mathematicians imagine that it is a fact of observation, and the observers that it is a theorem of mathematics. (Poincaré, 1892; cited from ESS, 1998)

In Section 4.5.5 we already saw that some scientists such as Bessel, Fechner, and Edgeworth dropped the assumption of symmetry (Steinbring, 1980). Pearson, in the late nineteenth century developed a class of distributions which were transformations of the normal distribution, and which could be applied to many more phenomena.
This section shows that the concept of distribution developed over several centuries. What is important for the historical and didactical phenomenology is that concepts such as distribution change over time. This means that we have to consider these concepts in a dynamic perspective, as Steinbring (1980) writes about the concept of chance:

> It is impossible to give a definition of chance that stays the same in all grades. This implies that the concept of chance should be carefully related to suitable contexts and extended by extending the contexts. This relation between foundation and application leads to a dynamic perspective of development with respect to the concept of chance. (p. 446; translation from German)

If we replace 'chance' by 'distribution', the quotation applies to our situation (in fact this holds for other statistical notions such as mean as well).

H22. A notion of distribution cannot stay the same in all grades. Accordingly, the representations in which students study distributions need not stay the same.

We allow students' informal and possibly sloppy characterizations of distribution in our teaching experiments and allow them to work with their own informal sketches before using more advanced representations and definitions.

4.7 Graphs

Graphs are crucial tools in statistical investigations, because we can see patterns and trends of frequency distributions that are hard to see from a table of numbers. In this section we ask ourselves from which phenomena statistical graphs were developed. Beniger and Robyn (1978) distinguish four problem areas in which the most important graph types were developed:

1 spatial organization (17th and 18th century), for instance Halley's map with lines of magnetic declination (1701);
2 discrete comparison (18th and early 19th century), for example Playfair's bar chart of import and export in Scotland (published in 1786);
3 continuous distribution (19th century) with histogram and ogive-shaped line graphs;
4 multivariate distribution and correlation (late 19th and early 20th century) with three-dimensional charts and correlation diagrams.

1. Spatial organization

Descartes (1596-1650) was convinced that imagination and visualization, and in particular the use of diagrams, had a crucial part to play in scientific investigation. One of his contributions, the coordinate system, still proves powerful today. A first major success of using coordinates in a Cartesian system was Halley's scatterplot of barometer readings against elevation above sea level (1701). This plot was an exception, though, because scientists had an obsession for tabular data (Beniger & Robyn, 1978). Between 1660 and 1800, even automatic graphs created by mechanical recorders to measure temperature, barometric readings, and tidal movements were routinely translated into tabular logs. Apparently, tables were considered clearer than graphs. It was not until the 1830s that scientific journals began to record graphs. We already know that students tend to focus on individual data values (2.2).

H23. Students initially tend to focus on tables and values. Even if they easily answer questions with the help of case-value plots, they still interpret these graphs as codifications of tables.

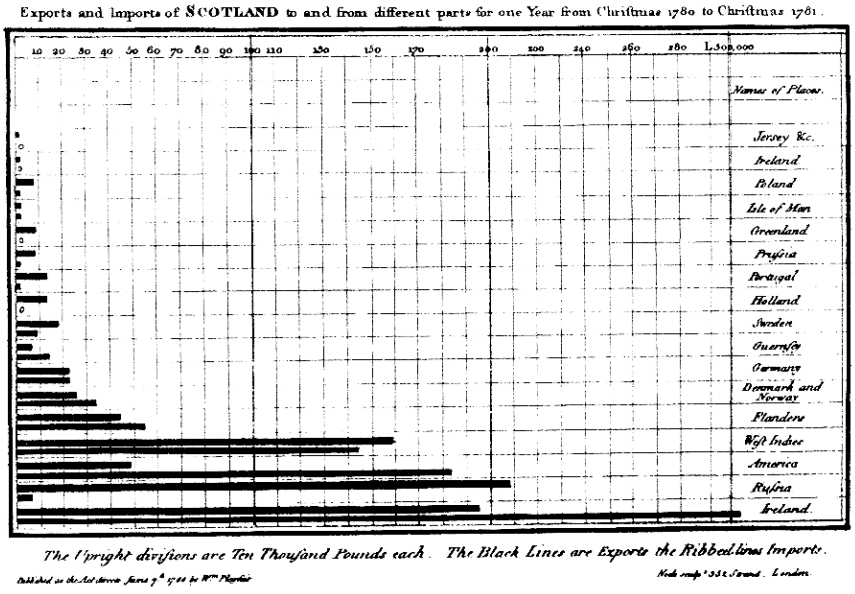


Figure 4.13: Playfair's bar chart of Scotland's import and export in 1781 (from Neeleman & Verhage, 1999, p. 21)

2. Discrete comparison

In 1765, Priestley published time-line charts with individual bars to compare the life-spans of about 2,000 celebrated persons who had lived between 1200 BC and 1750 AD. Not long after that, Playfair invented the first bar chart, which was published in 1786 and represented Scotland's imports and exports for seventeen countries in 1781 (Figure 4.13). Ironically, he made this graph due to a lack of data. Because he had no time series data, he graphed a single year as a series of 34 bars, and apologized to the reader for that. This graph can be considered a way to make a discrete quantitative comparison of import and export.
Playfair's bar chart was among the first graphs used and it resembles the representation of Minitool 1. Hence we conjecture:

H24. The bar chart in Minitool 1 is a representation that students easily come to understand.

This hypothesis was confirmed in the exploratory interviews and in the teaching experiments.

3. Continuous distribution

In the section on distribution we already demonstrated how several famous mathematicians had looked for functions that matched the distribution of error. Note that the graphs of Figures 4.9 and 4.10 were added for the modern reader as it was not until about 1820 that such graphs became more common. The problem of representing continuous distributions arose in vital statistics, the statistics of life information, and led to two important solutions: ogive-shaped line graphs and the histogram.
Fourier made a bar graph that represented the population of Paris by age groups and made a line graph of this, which led to the first appearance of a cumulative frequency distribution (1821). The first histogram was made by Guerry in 1833, who reorganized a bar graph to represent crime data that he had arranged in intervals of age and month. This led to a histogram. The term 'histogram' stems from Pearson (1895) (David, 1995; Schwartzman, 1994; Walker, 1931). Quetelet, then, was largely responsible for the further development of such graphs. In 1846, for instance, he published a symmetrical histogram with the curve of the normal distribution superimposed as the so-called 'curve of possibility'. The histogram emerged from organizing a bar graph. This could also indicate, from a historical perspective, that a histogram is a more advanced graph than a bar graph (see Section 2.2 and 2.3).

H25. The histogram is more difficult to learn than the bar graph.

Galton pictured the normal distribution differently from what we are used to nowa-

days (Figure 4.14 left). He wrote:

> I shall best explain my graphical method of expressing Distribution, which I like the more, the more I use it, and which I have latterly much developed, by showing how to determine the Grade of an individual among his fellows in respect to any particular faculty. (Galton, 1889b, p. 37) ['Grade' is a percentile; e.g., the first quartile is the 25th grade.]

This type of ogive-shaped graph was no exception in the beginning of the twentieth century, judging from Walker's remark in 1931 that such graphs were commonly used in school textbooks for statistics. Galton wrote that though the 'Curve of Frequency' (right) was generally used by statisticians, but then "turned at right angles," it was "far less convenient than that of Distribution [left]" (p. 49). He turned the frequency curve to "show more clearly its relation to the Curve of Distribution." But, he admits, "the Curve of Frequency has other uses, of which advantage will be taken later on" (p. 49).

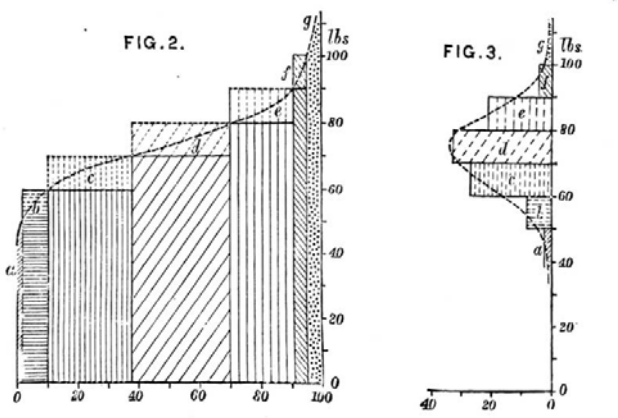


Figure 4.14: Galton's graphs of the normal distribution (1889, p. 38) on men's strength measured in lbs. Left is the 'Curve of Distribution' and right the 'Curve of Frequency'.

Galton called the type of graph on the left an 'ogive', after the architectural term (Bissell, 1996; Dictionary of Art, 1996). We have indeed found this ogive shape in many buildings in different countries (e.g. Figure 4.15).

H26. It is useful to let students deal with two representations of distributions similar to those of Galton, not just one (Figure 4.14).

Figure 4.15: Ogive shape in architecture (Kalamafka, Crete) and one of the ogive shapes from the Dictionary of Art (1996)

The reader may already have noticed the resemblance of Galton's graphs with Minitools 1 and 2, be they turned at an angle. Apparently, Galton found the cumulative curve (left in Figure 4.14) more useful for many of his purposes than the frequency curve (right), and he pointed at the transition between the two graphs that we also want students to see. It could indicate, but this is a tentative remark, that though the 'Curves of Frequency' are more common among statisticians, the 'Curve of Distribution' is easier for students to understand in some situations.

H27. The median is easier to conceive and develop in a representation that is similar to Galton's Curve of distribution than in a Curve of frequency; hence it may be easier in Minitool 1-type representations than in Minitool 2-type representations. It could well be important that the bars in Galton's distribution curve are vertical, because we conjecture that it is easier for students to read from left to right, and so to speak take the midrange, which happens to be the median, than reading from top to bottom, such as in Minitool 1.

We have not been able to test the hypotheses on the median in this study.

4. Multivariate distribution and correlation

By 1850, quantitative graphs had become accepted tools in statistics. The only graph of the problem area of multivariate data that we discuss is Galton's first correlation diagram, because we see an analogy with Minitool 3 as a sequel to Minitool 2. Galton's diagram, Figure 4.16, shows a bivariate distribution of head size and height. At the sides we see the univariate distributions of those variables. We interpret this example as supporting the idea of the Nashville team that a notion of univariate distribution is a prerequisite for really understanding bivariate distributions.

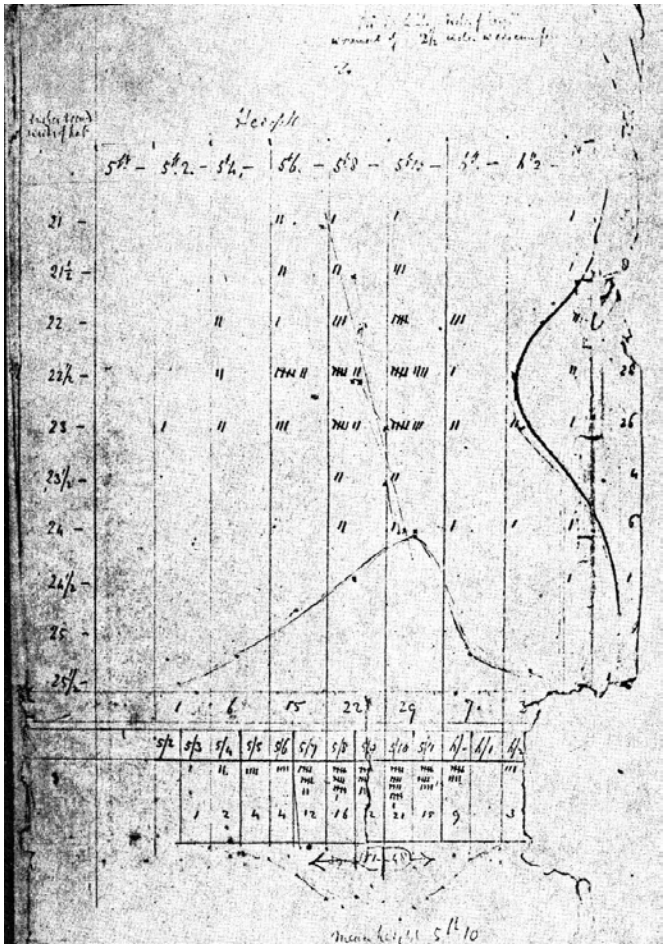


Figure 4.16: Galton's first correlation diagram on head size and height (Hilts, 1975)

We jump to the late 1960s for other major contributions to statistical graphing, namely in exploratory data analysis. The most famous newcomers are the stem-and-leaf plot and the box-and-whiskers plot, which were first presented by Tukey in 1969 as ways to display and explore data sets by hand. Because the basic box plot was often misinterpreted, Tukey and colleagues also invented alternative box plots with additional information (McGill, Tukey, & Larsen, 1978). From this we conclude that box plots are not that easy: even in the statistical world they sometimes led to confusion. Hence, they should also be handled with care in education.

H28. The box plot is one of the most advanced graph types used in middle schools. This is due to the incorporation of conceptual measures of center (median) and spread (quartiles).

We therefore propose to postpone the introduction of box plots until after middle school.
In addition to the four areas mentioned earlier, we cite Florence Nightingale (1820-1910) as a pioneer in graphical representations. Her main goal with those graphs was to convey to others the need to improve health care, for instance during the Crimean War (Cohen, 1984).

Table 4.1: Overview of the history of graphs, mainly from Beniger and Robyn (1978)

ca. 1350	Proto-bar graph of a theoretical function (Nicole Oresme)
17th cent.	Tables of empirical data (*Die Tabellen-Statistik* in Germany)
ca. 1660	Automatic recording device producing a graph of temperature (Christopher Wren)
1686	Edmund Halley's bivariate plot of barometric readings against altitude
1765	Measurement error as deviations from regular graphed line (Lambert)
1786	Playfair's bar chart
1801	Playfair's pie chart or circle graph
1821	Fourier's cumulative frequency curve of inhabitants of Paris by age groupings
1828	Mortality curve (Quetelet)
1830-35	Graphical analysis of natural phenomena appears in journals
1833	Guerry's first histogram of crime by age and months
1846	Quetelet represented urn schemata as symmetrical histograms with a "curve of possibility," later called normal curve
ca. 1855	Bar graphs and polar-area graphs on mortality by Florence Nightingale
1868	Statistical diagrams in a school textbook (Levasseur)
1874	Age pyramid, bilateral histogram (F. Walker)
1875	Galton's ogive graph of normal distribution
1884	Dot plot (see Wilkinson, 1999)
1969	Box plot and stem-and-leaf plot for EDA (Tukey)

We end this section with a chronological overview of types of graphs used in our research and graphs related to those (Table 4.1). We do not suggest that an instruction-

al sequence should follow this order. Taking Freudenthal's cue we try to understand the past well enough to design a "revised and improved version of the historical learning process" (4.1). What we see in history is that a lot of mathematics and statistics was already known before most graphical representations were invented. In education this is different: students know much less mathematics than scientists from around 1800 but they have encountered more graphical representations.

4.8 Summary

A historical phenomenology is an analysis of the development of concepts ('thought objects') in relation to the phenomena that gave rise to these concepts (2.1). The essential point of didactical phenomenology is to translate such phenomena into problem situations that are meaningful for students and create the need for organization by a particular concept. Knowing the historical development of certain concepts can help to anticipate a process of guided reinvention.

It can be demanding for instructional designers and teachers to put aside their knowledge of these concepts and take a student perspective. What may seem a minor step might have taken centuries to develop historically and might also be difficult to develop for students. A historical study can help to distinguish various aspects, problems, related notions and intermediate stages of the development of certain notions. In other words, it can help us look through the eyes of the students. This section is a summary of the results that turned out most useful for the didactical phenomenology and for the teaching experiments.

Estimation (H1, 7, 8)

Estimation of large numbers could well be one of the ancient origins of statistical methods. From a modern point of view we can recognize precursors to notions of average and sampling in those methods.

These thought objects of average and sample are used to handle the phenomenon of variability in what is estimated. This implies that we could use estimation tasks as the starting point of a statistics unit that supports a process of guided reinvention of average and sample.

Bars representation (H2)

Magnitudes can be represented by tallies and numbers, but also by the lengths of bars (Euclid, 1956). We assume that students can easily interpret bars as representations of data values such as in a value-bar graph, especially if the variable at issue has a time dimension (life span) or a one-dimensional physical connotation (wing span, height, braking distance).

Moreover, we assume that the bar representation of data values can help students in estimating means from data sets by using a compensation strategy. Better than from a table of values, they can see where the center of the data values is from a bar rep-

resentation such as a value-bar graph.

Midrange (H4)
Before the arithmetic mean was used to reduce measurement errors or to summarize data (16th century), the midrange was used for such purposes (9-11th century). From a modern perspective the midrange is not a very useful measure of center, because it is too sensitive to outliers, but with symmetrical distributions (such as most error distributions) this problem is less apparent. It is likely that students will also use the midrange as an initial way to find an average, for example when estimating total numbers. With skewed distributions students can then be challenged to scrutinize their midrange strategy. Similarly, skewed distributions can be used to create a need for a distinction between mean and median (H20).

Mean as an entity in itself (H6)
Considering the mean of a variable as a representation of a specific aspect of a population (percentage of dead letters, inclination to suicide) is much more recent than the other types of means we discussed. In the nineteenth century, the mean was used more and more as an entity in itself. This was especially apparent in cases where the mean did not refer to actual situations.
In 1877, for example, Peirce (CP 2.646) gave the example that in New York 14.72 persons lived in the average house. It is likely that students find this type of mean more difficult to understand than older types of means.

Sampling (H8, 10, 11, 12)
Just as there are different levels of using the mean, there are different levels of using sampling. In the estimation examples sampling was mostly implicit: the focus is on the total number and the sample helps to reduce variability in a smart way. In the Trial of the Pyx example of quality control of coins, the focus was on the weight and pureness of coins and sampling was necessary because not every coin could be melted to test its pureness. Both the total and the individual coins were clear units of thought.
Later in history, however, scientists became interested in more abstract aspects of populations, as the section on the mean as an entity in itself shows. More advanced methods of sampling became necessary, because measuring populations became too expensive or even impossible. In this case, a sample is a thought object with which something can be said about the population.

Measures of spread (H15)
One way to organize the variability of a data set is by summarizing it. This can be done with a measure of the spread. Historically the oldest measure of spread is the

range (David, 1998a). However, the range is sensitive to outliers; more robust measures are the interquartile range and the standard deviation.

Distribution (H21, 22)
A more sophisticated way to summarize or model a data set is by using the thought object of 'distribution'. In history, different distribution shapes have been proposed as summarizing the pattern in the variability of errors. Before the nineteenth century, distributions were assumed to be symmetrical. When statistics became used in more and more contexts including economics and social sciences, there was a need to make distinctions between different types of distributions (symmetrical or skewed, frequency or density distribution, sampling or population distribution).

Graphs (H23-25)
Before about 1800, scientists rarely used graphs because they preferred tabular data. Graphs are another way of summarizing data sets or patterns in variability. They can be used to represent distributions. What is interesting with respect to the normal distribution is that different representations were used; it was not only represented with the famous bell curve, but also as an ogive. This ogive shape also appears if vertical value bars are ordered by size, for example when students line up by height (Figure 7.19). It may be useful to use different representations of distributions to highlight different aspects of these distributions. By and large, the historical development of graphs is in line with the rationale of the Minitool representations.

On the basis of the present historical phenomenology, as well as prior research and exploratory interviews, a didactical phenomenology is formulated in the next chapter.

5 Exploratory interviews and a didactical phenomenology

As part of the preparation phase of the design research, this chapter presents the results of exploratory interviews and a didactical phenomenology of the key concepts of distribution, sampling, center, and graphs that can display distributions. One purpose of a didactical phenomenology is to find problem situations that can be used for the guided reinvention of the concepts, graphs, and types of reasoning that form the end goals. Such problem situations were found in three sources.

The first was the research presented in Chapter 2, the second the historical phenomenology of Chapter 4, and the third was a set of exploratory interviews. If we want to learn from prior research and from a historical phenomenology, we also need to know about students' prior knowledge. In what respect does the prior knowledge of Dutch students differ from that of the American students of the Nashville experiment? And what do students know that people in the past did not know?

To answer these questions, a set of exploratory interviews was used as the third source for the didactical phenomenology[15] presented in this chapter. Once we know more about the students' prior knowledge, we can identify possible starting points of a hypothetical learning trajectory (HLT). And once we know how the Dutch students react to certain activities that were used in Nashville, we will be better able to anticipate what will happen in the seventh-grade teaching experiments.

The following section describes the results of the exploratory interviews with 26 Dutch seventh graders. Together with the prior research (2.2 and 2.3) and the historical phenomenology, these results form the basis for a didactical phenomenology of the key concepts of statistics in this research. The last section of this chapter describes the first outline of an HLT for the seventh-grade teaching experiments. Schematically, this set-up can be represented as in Figure 5.1.

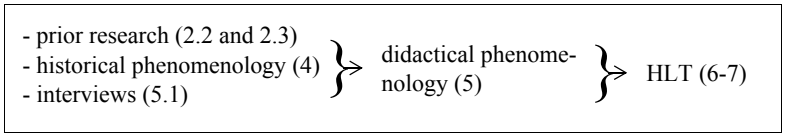


Figure 5.1: Schematic representation of the three sources of the didactical phenomenology, which in turn forms a basis for the HLT

15. For Freudenthal (1983a), a didactical phenomenology was carried out behind the desk, but we needed empirical input for writing this didactical phenomenology.

5.1 Exploratory interviews

5.1.1 Questions

To formulate an HLT, we needed to know more about the prior statistical knowledge of Dutch seventh graders: in particular, whether we could use the activities used in Nashville. As an indication of students' prior knowledge of statistics, we decided to investigate Dutch students' notion of average. Because multiplicative reasoning is important for statistical reasoning, we were also interested in how students would reason multiplicatively in statistical contexts. And the third sub-question was: How easily will the Dutch students solve the activities that were used in the Nashville teaching experiments? We chose two key problems from the Nashville research that had led to informal reasoning about distribution aspects and shape: a version of the battery problem with Minitool 1 and the speed trap problem with Minitool 2.

On the one hand, we expected Dutch students to be better able to understand the mean than the American students, because Dutch students learn to calculate their own grades for their reports with arithmetic means from grade 7 onwards (compared to American students who are used to a grading system with A, B, C,...).

On the other hand, we expected Dutch students might have difficulties in using particular graphical representations that American students were already used to, because Dutch students learn almost no statistics before grade 8, in contrast to American students.[16]

Table 5.1: Interview format

1	These horizontal bars represent the life spans of batteries of two brands (Figure 5.2). Which brand would you choose? Why? *(asked to fourteen students)*
2	At a certain spot in the city a lot of accidents happened. The police have placed a speed trap. Each dot represents the speed of a car (Figure 5.3). Below, you see how fast the cars drove before the speed trap was placed; above, after the speed trap was placed. Above 55 km/h drivers get fined. Did the action have effect? *(asked to fourteen students)*
3	What is the average? Can you estimate the average annual temperature from this table or graph? (Figure 5.4)
4	Can you calculate your report grade? Assume that three tests count twice and one counts once. *(For example, 7.8, 7.2, and 6.8 twice, and 8 once; a calculator was allowed)*
5	What does it mean that the average Dutchman watches television for 1.5 hours per day? Can you invent watching times that have 1.5 hours as an average?
6	What does it mean that the average family size is 2.5? Can you invent family sizes that have 2.5 as an average?

16. Compare, for instance, the key goals of the Dutch curriculum (Methodewijzer, 1998) with the Principles and Standards for School Mathematics in the United States (NCTM, 2000).

As described in 3.6, we interviewed 26 students, randomly chosen from seventh-grade classes of different levels, for about 15 minutes per pair. The interview questions are given in Table 5.1. For practical reasons, the graphs were presented on paper, not on a computer screen. In the following sections we present the results of these questions

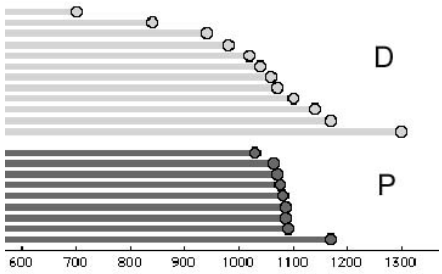


Figure 5.2: Battery problem in Minitool 1 (value-bar graph) as used for the interviews. The life span is in minutes. This data set is different from the one used in the teaching experiments. The samples of D and P are of different sizes (12 and 9).

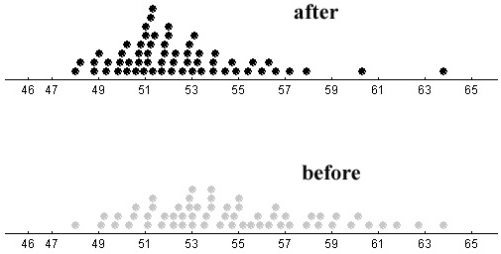


Figure 5.3: Speed trap problem in Minitool 2 (dot plot) as used in the interviews. The speed is in km/h. Above was the after situation and below the before situation. This is clumsy, but at that time we knew of no easy way to change the order. Both data sets consist of 60 values.

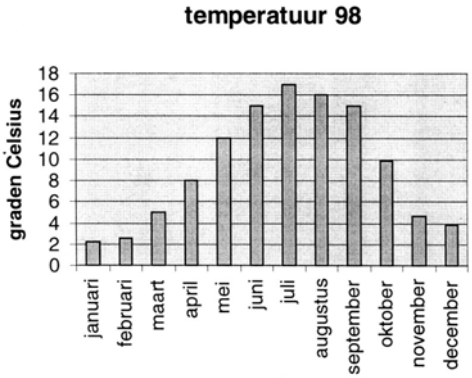


Figure 5.4: Average monthly temperatures in the Netherlands

5.1.2 Battery problem in Minitool 1 (question 1)

Seven student pairs solved the battery problem, and seven pairs solved the speed trap problem; one pair did both because there was time left. No student found it difficult to read off the values from a value-bar graph, but all of the students first chose the longest bar of brand D: "You can just take the longest, can't you?" This shows that these Dutch students also tended to focus on individual data values (cf. Section 2.2). In terms of a didactical phenomenology, they initially organized this problem situation by localizing individual data values and not by conceiving the general shape or by regarding the data set as a sample of one battery brand.

When we explained that the longest bar only referred to one battery that lasted a bit longer than the rest of that brand, students started to look at various other aspects, for example:

1. You have to be lucky.
2. It depends on which (battery) you buy.
3. Those four (of D) are better than those (large group of P).
4. Likelihood that you get a short one with this brand (D).
5. P has more of the same value.
6. With P you know how long it lasts.
7. The P ones are closer together, they're all even, not "which one shall I take."

These examples vary in quality. The first two arguments probably show a case-oriented view, because students focus on the single battery that might be bought. The student who used the third argument paid attention to a specific category within the data sets, a categorical view, which we consider in between a case-oriented and an aggregate view on the data sets. Arguments 4 to 7 indicate that students looked at the whole data set. In argument 4, the "likelihood" probably refers to a relatively large part of low values in the subset of D compared to that of P. Arguments 5 to 7 suggest a sense of consistency, though the focus also seems to be on the battery that is going to be bought.[17]

As these examples illustrate, the battery activity could function as a springboard for discussion about majority, outliers, chance, and predictability. It would be necessary to support students in making their fuzzy and informal notions more exact, for instance by quantifying expressions such as 'close together' by a measure of spread, and 'how long it lasts' by a measure of center. Somehow, we would have to prevent too many students from only looking at one value (see Section 6.4). Moreover, we decided to use a different data set, so that students would first compare two subsets with the same number of data values (in this data set, there were 12 data values of D and 9 of P). We expected this to be easier as a starting point.

17. This is similar to the so-called outcome approach (Konold, 1989), which is common in probability contexts.

5.1.3 Speed trap problem in Minitool 2 (question 2)

Ten of the fourteen students to whom we presented the speed trap problem were able to read the dot plot without further explanation. The other four found it hard to read the dot plot in the first instance, but quickly came to understand it. One student initially thought that the cars ran into each other. From what she said, we concluded that she interpreted the dots as cars that were literally close to each other, as on a map or in a picture. Apparently, she did not immediately interpret the dots as signifying the speeds of cars.

The majority of the students quickly answered that the speed trap had an effect, albeit "not a very large effect." A common argument was that fewer people drove more than 55 km per hour after the speed trap had been installed (55 km/h was the speed at which people would be fined). Because both conditions consisted of the same number of measurements, this was a valid argument.

What bothered us about the context was that students used 55 as a cutting point and did not argue about a hill or a majority. For the HLT we decided to change the formulation of the problem and leave out the number 55, in the hope that students would look at the whole distribution and not cut it into two parts. Thus we tried to avoid the distracting influence of focusing on multiplicative reasoning instead of the whole distribution (R8).

One issue that struck us was that several students gave commonsensical answers that were often not based on data, but on what they knew about the context. For example:

> People who know will drive slower.
> It depends on the day too; on Sunday people may drive more slower.
> It helps because above 55 km/h they get fined.
> They get a fine so they don't drive too fast anymore.
> I think the people knew there was a speed trap and slowed down.

This is not really surprising, because these students had no prior experience with data analysis. For the HLT we concluded that the teacher should establish the norm of needing to reason with data when available.

We also wanted to have an indication of how well Dutch seventh graders reason multiplicatively. We therefore asked ten students to reason from new hypothetical data (Table 5.2):

> Assume there were six cars below 55 and six above in the before situation versus twelve below and six above in the after situation. What would you say then?

Table 5.2: hypothetical situation of speed problem

speed trap	<55 km/h	55+
before	6	6
after	12	6

Students' skills in multiplicative reasoning varied. Some said that there was no improvement because there were still six cars driving too fast; they reasoned additively, but within the context this is not really surprising. Maybe they thought that for safety the absolute number of cars driving too fast is more important than the percentage. One girl was in doubt:

> The number is important, but also the whole. You cannot really tell. There are also more cars here (in the after situation).

One boy noticed that "it had effect in percentages."
Students' skills in calculating percentages or proportions also varied considerably. Two of them were not able to calculate the percentages of cars driving too fast; two answered the question with informal terms such as 'the most;' most of them needed a little help, and four students easily found either 50% and 67% or $\frac{1}{2}$ and $\frac{2}{3}$.
The speed trap problem turned out to be a rich problem with possibilities for talking about shape ("the hill is more compact here"), and for extending it to more difficult situations that ask for multiplicative reasoning. As with the battery problem, we concluded that the speed trap would be a useful activity in the HLT with a potential for progressive mathematization. The difference between the before and after situations might be quantified with the mean or median, and the difference in spread might be quantified with quartiles.
Because the battery problem preceded the speed trap and Minitool 1 preceded Minitool 2 in the Nashville sequence, we had expected that students would find the battery problem easier than the speed trap, and Minitool 1 easier to use than Minitool 2. In contrast to that expectation, the students needed a little more time and help to answer the battery problem than to answer the speed trap problem. Reading the two graphs was not a problem except for four students reading Minitool 2. We then assumed that if we spent more time talking through the data creation process in the battery context, the difference in difficulty between the two problems would diminish. Additionally, we decided, in the teaching experiments, to pay attention to whether these students found Minitool 1 easier to use for analyzing data sets than Minitool 2, because insight into this issue would be relevant for the HLT. We return to this issue in Section 6.13.

5.1.4 What is the average? (question 3)

With respect to the question on the average we need to mention that the Dutch word for average, *gemiddelde,* refers to both the informal meaning of average and the statistical meaning of arithmetic mean, but it does not function as a collective noun for mean, median, and mode as the term 'average' sometimes does in English. In answer to this third question on the average, 13 of the 26 students mentioned the algorithm or parts of it, and ten mentioned other aspects such as 'most', 'about', 'roughly', 'in between', 'a bit in balance', and 'midpoint' (Table 5.3). In the contexts of other in-

terview questions, students also said things such as "a large amount" or "half is below, half is above" when looking for the average.

Table 5.3: Students' reactions to the question, What is the average?

What is the average?	freq	Possible statistical interpretations and aspects
The half. The whole, and in between the half, that is the average	1	Part of the algorithm: dividing by two
Everything together	1	Part of the algorithm: adding all values
Add and divide by 2	2	Simple algorithm
Add and divide by the number	9	Algorithm
The most	2	Mode, typical
What you think it is roughly	3	Estimation, representativeness, true value?
The mean is about a bit in balance	1	Balance point
In between ("15 is halfway 10 and 20"); what is in the middle	3	Midrange
The midpoint	1	Midrange, median, center of gravity
A large amount (context of TV watching)	2	Majority
It is put in balance (context of TV watching and temperature)	1	Balance point
Half is above, half is below (temperature context)	1	Median
Compensation strategy in bar graphs (context of battery and temperature)	5	Mean

Five students (both at the *vwo* and *mavo* levels) reinvented a visual compensation strategy–two with the battery graph from question 1 (Figure 5.2) and three with the vertical temperature bar graph (Figure 5.4). One boy said:

> I take off everything that is above 8 (degrees Celsius) and add that to what is under 8; you do that by filling up till 8 (later he agreed with his peer that 9 or 10 ^{o}C would be a better estimation).

When we asked other students who did not spontaneously use that compensation strategy what they thought of estimating the mean in the bar graph by cutting and pasting, most seemed to understand it, but one boy opposed.

> It is not possible, because in July it is warmer than in January, so it is false.

Thinking of the literature on the mean and the Nashville experiment (R9), we found

it promising that almost half of the students did not immediately mention the algorithm or parts of it, but mentioned qualitative aspects such as midpoint, the most, or "what is in the middle." There was no indication that students calculated the mean for whatever statistical problem they encountered. We concluded that the drill and practice problem that many American researchers have observed (2.2) did not apply to the Dutch situation. Hence, we did not need to follow the attempts of the Nashville team to avoid the mean for this reason (P7). In Section 5.3 we further elaborate on the position of the mean in the HLT.

5.1.5 Weighted mean (question 4)

Question 4 asked students to calculate their own report grade, which was a weighted mean. Of the 19 students asked, two did not do it correctly, nine first divided by 4, realized that something was wrong, and changed to dividing by 7; the remaining eight students found the right answer immediately. If we take into account that even college students have trouble with weighted means (Hardiman et al., 1984; Pollatsek et al., 1981), we can conclude that these students were reasonably skilled in calculating means in the context of their report grades. However, it is likely that their performance of weighted means in other, less-known contexts would not be as good.

5.1.6 Average Dutchman (question 5)

Students reacted differently to the question of what an average Dutchman watching 1.5 hours of television daily means. Nine of the fifteen students were not surprised by the notion of an average Dutchman, but six found it strange. A few students said that the average Dutchman did not exist; one called it a typical Dutchman; two students explained it as "a large part of the Dutch"; one as "it varies around it." Two students, Leila and Jenny, thought that an average Dutchman would be half a Dutchman (they also thought that the average was "the half"; see Table 5.3) and their halving strategy was persistent:

Interviewer:	Roughly how many hours of TV do you watch per day?
L.:	Ten or so?
Interv.:	Per day? [surprised]
J.:	So much?
Interv.:	Isn't that impossible?
L.:	Hm. [thinking]
Interv.:	And you? [to J.]
J.:	Five hours or so.
Interv.:	Per day? Each day? And how much on average? [just checking]
J.:	2.5.
L.:	2.5. [They divide by 2 again.]

Of the nine students we asked to invent watching times that would yield 1.5 as an average, two found this very hard even with help, four students succeeded with supportive questions, and three succeeded without any help and mentioned for example

1 and 2 hours as possible watching times. We concluded that students' notions of the average Dutchman varied considerably. We also got the impression that their skills in calculating averages seemed to be separated from their informal knowledge (as expressed in "what it is roughly").
Historically, the average man was an invention by Quetelet (nineteenth century). Though the idea of an average man is culturally quite accepted now, it is quite a different matter to have an informal image of this idea than to understand its meaning in terms of calculations, as is shown by the next section as well. This contrast raised the question of how students' informal ideas about average could be linked to their idea of the mean as an algorithm.

5.1.7 Average family size (question 6)

Although most students had sensibly answered the previous question about watching television, eleven were confused by the question of what was meant by an average family size of 2.5. Only five understood it. Seven students thought the 'point 5' would be a child. Two said that half people do not exist and one joked that "the point 5 did not have legs." Apparently, the students experienced the two contexts of the 1.5 TV hours and the 2.5 people very differently. This is not really surprising, because half hours are common and half people are not.
For the didactical phenomenology this means that it is difficult for students to deal with the mean as a construct in cases in which the resulting number cannot correspond to an existing situation. From the historical phenomenology, we had expected that the mean as an entity in itself representing aspects of populations would be the most difficult face of the mean (4.3.4). From the interviews we concluded that a further distinction into discretion (e.g. family size) or continuity (e.g. watching hours) was necessary.

5.1.8 Conclusions from interviews

What can we learn from the interviews for the didactical phenomenology? These seventh-grade students had already learned about the arithmetic mean, but our impression was that their informal understanding of the mean and their knowledge of the algorithm should somehow be linked better. Their understanding of the mean turned out to be better than the students in Nashville (McGatha et al., 2002), and the Dutch students were certainly not drilled to calculate means for solving statistical problems.
For the beginning of the HLT, we searched for suitable problem situations to refine students' intuitions of the mean and link that intuition to their knowledge of the algorithm of calculating the mean. The battery and speed trap activities, which were designed for the Nashville experiments, were neither too easy nor too difficult for the Dutch students. We concluded that we could use these activities in the HLT for the Dutch seventh-grade students as well. We expected that the Dutch students

would need less time than the Nashville students to solve those problems, because we would only work with *havo* and *vwo* students, that is with the students being prepared for higher vocational education and university (about 35% of all students follow these tracks of education).

5.2 Didactical phenomenology of distribution

Because the concept of distribution is central to the research, we present a didactical phenomenology of distribution in relation to other key concepts such as sampling. In the next section we then turn to specific aspects of distribution: center in particular. Note that the ideas on distribution as presented below are more explicit than they were before the teaching experiments.

The basic phenomena that statistics is about are uncertainty and variability. To detect patterns in the variability we can take a sample, do measurements, and thus create data. If a data set is created, it can be analyzed for patterns and trends by using suitable diagrams. A key concept in this analysis process is the concept of distribution, which can be seen as a pattern in variability. Distribution has various aspects such as center and spread, but also density, skewness, kurtosis, et cetera. Table 5.4 visualizes the relations between those statistical ideas.

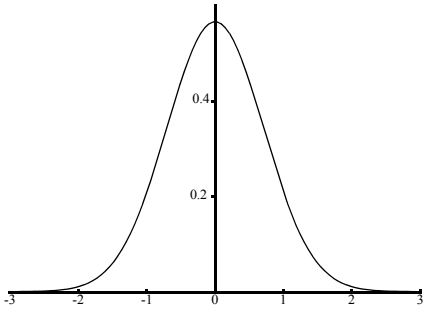
Table 5.4: Structure of distribution in relation to other key concepts of statistics

distribution (conceptual unity, pattern in variability; represented in diagram)			
center (mean, median, midrange)	**spread** (range, IQR, SD)	**density**	**skewness** etc.
data (empirical plurality; represented in a table)			
sampling (measurement leads to data)			
uncertainty <> variability (lead to necessity of sampling and data)			

The power of statistical data analysis lies in describing and predicting aggregate features of data sets that cannot be noted from individual cases. Consequently, aggregates form an essential topic in this didactical phenomenology. As mentioned in Chapter 2, students tend to conceive a data set as a collection of *individual values* instead of an *aggregate* that has certain properties (Ben-Zvi & Arcavi, 2001; Hancock, Kaput, & Goldsmith, 1992; Konold & Higgins, 2002). An underlying problem is that middle-grade students generally do not see 'five feet' as a value of the variable 'height', but as a personal characteristic of, say, Katie. In addition to this case-ori-

ented view, students should learn to disconnect the measurement value from the object or person measured, and consider data against a background of possible measurement values. The importance of developing a notion of distribution is that it is an organizing conceptual structure with which students can conceive the aggregate instead of just the individual values (Cobb, 1999; Gravemeijer, 1999b, c; Petrosino, Lehrer, & Schauble, 2003). In other words, with a notion of distribution and with the help of suitable diagrams, students can learn to conceive patterns in data sets and develop an aggregate view of data. To help students come to view data sets as objects with aggregate characteristics, we should therefore pose problems that can be solved by reasoning about the characteristics of data sets and not by looking at individual data points.

Before the teaching experiments in grade 7 started, one of the few things we explicitly wanted to accomplish was that students would come to see mean and median as characteristics of a distribution and not just as operations on data values. For example, when seen as group descriptors, mean and median can be used to compare two data sets. As Zawojewski and Shaughnessy (2000) note, students can only sensibly choose between these measures of center if they have some notion of distribution. But what do we mean by 'distribution' for students with hardly any statistical background? It is clear that one cannot teach them the probability density function of the normal distribution (Figure 5.5), because that would be too formal. We decided that 'distribution' had to mean a frequency distribution, at least in the beginning. Nevertheless we also decided that density had to be addressed in an informal sense, as it had been in the Nashville research (2.3).



$$f(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$$

Figure 5.5: A graph and a formula of the probability density function of the normal distribution

Our aim with the instructional unit was that students would come to see center, spread, and skewness, as characteristics of a distribution. Mean and median are then measures of center (of the distribution); the range and standard deviation (not used in the present study) can become measures of spread (of the distribution); skewness

can be characterized informally by where the majority of the data values are in relation to the extreme values.
To clarify this aim as opposed to statistics as 'number crunching' (2.2), we came to distinguish an upward and downward perspective with reference to Table 5.4. In the upward perspective, students tend to operate on individual data values, such as calculating the mean or quartiles. In the downward perspective, students use a notion of distribution with its characteristics to model data. As the research literature indicates (2.2), novices in statistics generally have an upward perspective only. Experts can combine the two perspectives: they can regard a mean, for instance, as a calculation on data values but also interpret it as an estimator of the theoretical mean of a distribution, that is as a characteristic of that distribution.
It is clear that middle school students cannot reach the full downward perspective. However, in our research we have become convinced that we should and could make progress in that direction. In other words, students should learn to model data at an informal level with an informal notion of distribution and come to see measures of center and spread as characteristics of a distribution (pre-stages to the downward perspective).
Our aim was not to work through an upward perspective in understanding of frequency distributions, and only then to let students develop a downward perspective. Instead, we wanted students to develop a downward perspective from an early stage. The question is how, and the answer lies, as R14 of the Nashville team states, in the notion of *shape*. It is possible to infer characteristics of a distribution from its shape in a particular graphical representation. 'Shape' offers a way to think and talk about the characteristics of a distribution, whether a frequency distribution or a density function. For example, we expected that students, as in the Nashville research, would learn to conceive the distribution as a whole with the help of 'hills'. Of course, the shape of a distribution depends on the graph type used; a normal distribution in a value-bar graph has a different shape than in a dot plot (cf. Figure 4.14).
Dutch students are acquainted with graphical representations such as the bar graph and the Cartesian coordinate system. The interviews indicate that the students, in general, did not find it difficult to interpret the value-bar graph of Minitool 1 or the dot plot of Minitool 2. As motivated in Section 2.2, we assumed that case-value plots (such as value-bar graph and dot plot) would be easier to use for students than aggregate plots (such as histogram and box plot). We decided to follow the rationale of the Minitools (2.3.2) as long as empirical observations did not indicate another direction. As mentioned in Section 2.3, Cobb (1999) describes how a seventh-grade class proceeded from the practice of dealing with collections of data to dealing with distributions of data. We decided to try and replicate that, albeit with a larger role for the mean and for sampling.

Reification of distribution

As stated above, one of the goals of the HLT was that students would come to conceive a distribution or its shape as an object that has certain characteristics. The question of how this process evolves is part of the second research question. To answer the question of how distribution can become an object, we could not build upon the Nashville research. Instead we tried to apply different theories on the formation of mathematical objects.

– Tall and colleagues (2000) use the notion of 'procept' to refer to the two-sidedness of mathematical objects as procedures and concepts.
– Dubinsky (1991) uses the notion of 'encapsulation' to describe object formation. A step-by-step action (A) becomes conceptualized as a process (P), which then is encapsulated as a mental object (O). This object can then become part of a schema (S). This four-part theory is referred to with the acronym APOS.
– Sfard (1991) describes the two sides of mathematical concepts as operational and structural, and stresses the importance of discourse in the process of object formation (Sfard, 2000a, 2000b). The last step of this object formation process is called 'reification'.

A common element of those theories is that procedures can become an object (e.g. Tall et al., 2000; Van Oers, 2000). For example, the expression *2x+6* can be interpreted as a procedure: multiply a number *x* by 2 and add 6. Sometimes, however, it is necessary to interpret this expression as an object, for example, when substituting this expression in equations, such as $\frac{12}{2x+6} = 3$. To solve this problem it helps to regard the denominator as an object, the number 4, because otherwise it would probably be a trial and error procedure of trying out numbers for *x* (see Drijvers, 2003).

The focus of these modern theories is on the concepts of number and function. If a set of objects is counted (a procedure), the result is a number (an object). The function as a calculation procedure leads to an outcome, and as an object, it can be manipulated (e.g. substitution, multiplication, transformation) and described (e.g. linear, quadratic). When we tried to apply these theories, we ran into the question of what would be the operational side of the concept of distribution. The only thing we could think of was the process of sampling. In Section 9.7, we describe in what sense students might interpret distribution as a procedure.

More than with functions, the concept of distribution shows similarities with composite units (Schifter & Fosnot, 1993). The number 10, for instance, can be seen as a collection of ten individual things, but also as a unit. For young students, it is often difficult to combine these two views as a composite unit. Similarly, students find it difficult to see a data set as a unit composed of individual cases.

Sfard's theory offers a heuristic for object formation entailing that one should look for a situation in which students need a process as an object at a higher level. For instance, in the preceding example, $\frac{12}{2x+6} = 3$, it is easier to solve the equation

when taking the denominator $2x + 6$ as an object than taking it as a procedure (cf. Drijvers, 2003). In Section 6.11 we describe how we used this idea of manipulating distributions as objects to create the need of distribution as an object.
There is yet another important aspect in the theories on reification which has to do with using predicates for objects. Peirce, Piaget, and Dienes, for instance, described how a predicate of a subject can become the subject of a further predicate (see also Van Oers, 2000, on predication). We envisioned that a mean, spread, and other notions should become characteristics of a distribution, or predicates of an object, but we did not yet know how. This topic is addressed in Chapters 8 and 9.

5.3 Didactical phenomenology of center, spread, and sampling

In this section, we focus on center but also treat spread and sampling. Traditionally, measures of center such as mean and median are taught before distributions come into play. However, students need to have a sense of distribution before they can sensibly choose between such measures (Zawojewski & Shaughnessy, 2000). For example, if there are outliers or if the distribution is skewed, we might prefer the median. Furthermore, students must have a sense of center before they can interpret a mean or median as a measure of center or central tendency. This leads to the question: which should be developed first? In our view, students need to develop notions of center and distribution concurrently. And this, in turn, means that we need to define levels of understanding center and distribution.
To map out an HLT for center in relation to distribution, we need to know about the notions students already have and to make conjectures about which types of center can be developed in which contexts and in which order. To do so, we go back to the exploratory interviews and the historical study.
It was not always clear how to interpret students' answers to the question of what the average was. When one boy said "the midpoint," did he refer to the point in the middle of the lowest and highest value, to the middle-most value, or to a center of gravity? We were not sure. At least we can conclude that the student answers together cover many aspects of the average. Additionally, we conclude that the students made no clear distinctions between the different aspects of the average values.
Similarly, the historical examples on estimation illustrate that it can be difficult to make implicit aspects of average values explicit. Did Thucydides really think of the midrange in estimation example 3 in Section 4.3.1? When people organized their world and solved problems, they were urged to become clearer and define their methods more precisely than before. The etymology of the word 'definition' reflects this as Latin 'finis' means end, border, or boundary. "When you define something you 'put boundaries around' what it can mean. A good definition puts an end to confusion about what a term means" (Schwartzman, 1994, p. 68). Note that putting an end to confusion is mostly temporary, both in history and education. In line with the

RME tenet on phenomenological exploration and the heuristic of guided reinvention (2.1), we strove for a learning process in which students would first explore the subject area before they understood and appreciated clear and more formal definitions. They should first have an image before they sharpen it. This resonates with the pedagogy mentioned in Section 2.2, but most textbooks take the opposite direction. They define mean, median, and mode, and then let students practice the procedures and applications.

From the research literature, it is well noted that mean and median as representative values are hard to develop for students (Section 2.2). Mokros and Russell advise postponing the algorithmic aspects of the mean until late in the middle grades, "well after students have developed a strong foundation of the idea of representativeness" (Mokros & Russell, 1995, p. 38). Representativeness is important because the average as a representative value has to represent the total, that is the data set or distribution, for example when comparing two situations.

We could not follow the advice of Mokros and Russell for the mean, because Dutch students in grade 7 already know the algorithm of the mean. However, we could try to create opportunities in which students could link that algorithmic knowledge with the mean and median as representative values. We used the historical phenomenology as a source of inspiration while keeping in mind that these students had already learned the algorithm. From the history we identified three different levels of representativeness in using the mean:

1 *Estimation*. For example in the case of the number of leaves and fruit, the average was the number of leaves on a typical branch. In the estimation examples (4.3.1), the mean was used as a multiplicand to find a total. The role of the mean as a representative value is rather implicit in such cases. The mean should not be too little and not too much, because otherwise the total would be too small or too large.
2 *True value*. The mean as an estimator of a true value so to speak represents that true value (4.3.3). Konold and Pollatsek (2002) characterize true values as signals in noise, a metaphor that probably stems from telegraphy.
3 *Entity in itself.* The mean as an entity in itself represents an aspect of a population (4.3.4). The mean could be something that does not exist, such as an average family size of 2.5.

From the history of the mean we conjectured that the first level would be easier for students than the second and we concluded that there was an important conceptual leap from the first and second levels to the third. This could imply that the latter type of mean would be more difficult for students than the first and second types of mean, and this was indeed reflected in the exploratory interviews. When we asked students to explain how families could have an average size of 2.5 people, several students thought that this referred to two adults and one child. We decided not to address this issue in our HLT, but to make sure that students first linked their qualitative insights

about the average with their algorithmic skills. We present five approaches of how this could be done while also drawing students' attention to how data values are distributed. The historical phenomenology served as a source of inspiration formulating those approaches.

a To estimate a large number, a mean value can be multiplied by the number of cases ($n \cdot \bar{x} = \Sigma$). The kind of average value used might initially be implicit because the focus is on finding the total. In this way, we expected that students, while organizing the problem situation, would develop qualitative aspects of the mean such as representativeness, intermediacy, balance, and compensation (cf. Mokros & Russell, 1995; Strauss & Bichler, 1988). We searched for pictures of demonstrating people, stars in the sky, flocks of birds, and ended up with a picture of a herd of elephants (Boswinkel et al., 1997) from which we decided to ask students to estimate the total number. These problem situations 'beg to be organized' by a notion of average.

b The first variation on the theme of estimation is when we have or estimate a total and use the mean value to find a number ($\Sigma / \bar{x} = n$). An example of this is the so-called 'polar bear problem', designed by Van den Heuvel-Panhuizen (1993, 1996):

> How many students weigh as much as one polar bear?

Nelissen (1997, 1999) describes how third-grade students guessed that a polar bear would be 500 kg and an 'example child' would be about 25 kg, which resulted in 20 students, but not all students bought into this context of the polar bear. Inspired by this activity we thought of an elevator context, but the teacher we worked with in grade 7 proposed to change the context:

> How many 12-year-old students could go into the basket of a hot air balloon if normally eight adults are allowed?[18]

We expected students to use informal sampling methods to find an average value (see Section 6.3). Discussions about the reliability of the methods might bring up many issues of sampling and distribution (representativeness, sample size, sample method, shape of the distribution).

c Fair share (H3) has to do with finding the mean from a total and a number ($\Sigma / n = \bar{x}$). This calculation answers the question of how much everyone would get after fair redistribution or reallocation (4.3.3). The fairness origin of average in combination with children's intuitions on fairness implies that fair share might be a suitable instructional context to develop an understanding of the mean as well (e.g. Boswinkel et al., 1997). Related to this fair share is the mean as a mea-

18. The eight adults do not include the driver.

sure for fair comparison, for instance if we need to compensate for the number of values in different groups. We then use parts per million, a percentage, gross national product per capita, et cetera. Cortina and colleagues (1999) call this type of mean the "mean as a measure." The historical examples of coin testing (Section 4.4.3) and of measuring a 16-foot rod were also means as measures (Section 4.4). A disadvantage of a fair share approach in education is that it might stress the computational face of the mean.

d Another way to stimulate a more qualitative notion of average was inspired by the compensation strategy that five students in the interviews used and by the Greek way of representing magnitudes. The Greeks used bars to represent magnitudes of different kinds (4.3.2), but generalization of the mean to *n* values had to wait until the sixteenth century (4.3.3). In fostering a visual estimation of the mean in case-value bar graphs, we expected to benefit both from the Greek representation and the generalized definition (H2). On the basis of this historical study we conjectured that students would not reinvent a compensation strategy when using dot plots or balance models.

e Another possible way to foster a qualitative view on average and a sense of how data values are distributed is by allowing the midrange as a first, rough way for students to organize center. Historically, the midrange was one of the predecessors of the mean. In other words, the middle of the range was used as an indication of center or a true value. Hence, students might also structure problem situations by using that value (H4). In most school textbooks, the midrange is avoided because it is sensitive to outliers. Without the historical study, we would probably not have thought of the midrange as a precursor to the mean or of allowing the midrange as an initial measure of center.

For understanding the average values, it is also necessary to learn and examine how data values are distributed. If data values are distributed symmetrically, there is no evident need to distinguish the midrange, median, or mean. A skewed distribution can show the limitation of the midrange as an organizing measure of center, and it could make the differences between midrange, mean and median a topic of discussion in relation to outliers and the skewness of data (H20). If we want students to understand the drawbacks of this measure, we should use skewed data at some point in the HLT.

So far, we have not addressed the median. Before the teaching experiments we had no clear image of what to do with it, but we knew that the median's historical development was a kind of side branch of the mean's. The Nashville research shows that it is difficult to let students develop the median as a representative value.

We decided that the median should be dealt with in the HLT, because it seems easy to calculate and it is useful as a measure of spread in combination with the quartiles,

but we did not yet know how.[19] We decided to start with the mean, because students were already familiar with it and because we saw approaches (*a* to *e*) of letting students develop the mean as a measure of center.
With the estimation activities we can clarify what we mean by a didactical phenomenology. We were in search of problem situations in which students would be challenged to reinvent methods or thought objects to organize certain phenomena (2.1.3). We expected students to use intuitive notions of average, density, and sampling to estimate the total numbers of elephants and the number of students in the balloon basket. Additionally, we envisioned that in discussing these issues that students would think more explicitly about center and sampling, and perhaps even the distribution of elephants and student weights.
The image that emerged from the historical and didactical phenomenology is that the mean is a rich notion with many faces. Because we did not want to stress the algorithmic face of the mean in the HLT, we decided to focus initially on the first two types of estimation (*a* and *b*), and not on fair share (*c*).
As mentioned earlier, there was little research on students' development of variation or sampling when we started this research. We summarize a few of our thoughts. The most basic measure of variation, both historically (David, 1998a) and didactically, is the range. We also saw historical support for the conjecture that the interquartile range is intuitively clearer for students than the standard deviation (H17; cf P8). Understanding of the interquartile range could be prepared with the four-equal-groups option in Minitool 2.
On the basis of the historical phenomenology we assumed that some instances of sampling would be easier to understand for students than others (H10, H11).

1 A situation such as the trial of the Pyx, in which the unit of thought is a single item (for instance, a coin), is probably easier for students to develop a notion of sampling from than
2 a situation of stratified sampling or
3 random sampling in which the focus is on an aspect of a population.

One reason we assume that stratified sampling is easier for students is that it suggests that we can deliberately create a representative sample. We decided to use the first situation as an initial attempt to address sampling (6.8).

19. Please recall that the historical phenomenology of the median was only carried out after the seventh-grade teaching experiments.

5.4 Initial outline of a hypothetical learning trajectory

The main point of departure was that students would develop notions of variability, sampling, data, and distribution in a coherent way (2.4). In particular, we strove for the reification of the notion of distribution by making the shape of distributions a topic of discussion. On the basis of the historical and didactical phenomenology, we thought of using estimation tasks such as the elephant and balloon activities as a starting point of the HLT. We conjectured that students would use a notion of average to find a total number and would look at the way the elephants were distributed in the picture. During the balloon activity we expected students to use a notion of average as well, and possibly informal notions of representativeness and sampling. Variability in those contexts would be no problem: the elephants were not uniformly distributed over the picture, and students know that not every student has the same weight. As written earlier (2.4), we decided to pay more attention to sampling than the Nashville team had, but apart from the elephant and balloon activities we did not yet know how.

By and large, we followed the Nashville rationale of the Minitools (2.3.2). Students would for instance first solve the battery problem with value-bar graphs (5.1.2), which would presumably lead to similar reasoning about aspects of distribution such as majority, consistency, and outliers, as the Nashville team had reported.

Furthermore, we expected that students would come to reason with hills when working on the speed trap problem (5.1.3), and that they would explore ways of describing the speed trap distributions with the grouping options of Minitool 2, especially four equal groups and equal interval width. We strove to foster a shift from qualitative to quantitative reasoning, that is from reasoning about hills to classes, quartiles, or percentages. We also strove to end up with the box plot and histogram representations to describe and represent distributions in conventional ways. A possible endpoint of the HLT was that students would assume a stability of shape and that they would be able to interpret shapes in the context and, vice versa, to relate changes in a context to changes in the shape of a distribution. In short, this initial outline can be characterized as challenging students to reason about aspects of distributions in increasingly sophisticated ways with increasingly sophisticated notions and graphical representations.

6 Designing a hypothetical learning trajectory for grade 7

Chapters 4 and 5 describe the preparation phase of the research. Chapters 6, 7, and 8 concern the other phases of the design research cycles in grade 7.

- Chapter 6 focuses on the *design* process of a hypothetical learning trajectory (HLT) in three teaching experiments.
- Chapter 7 focuses on testing the developed HLT during the last *teaching experiment* in grade 7.
- Chapter 8 is a *retrospective analysis* of the symbolizing process when students learn to reason about the shape of a distribution.

In the present chapter, we discuss issues relevant to the hypothetical learning trajectory (HLT) or, more generally, to the evolving instruction theory for early statistics education such as the importance of talking through the data creation process and students' roles as data analysts. We further describe the design process of instructional activities in three cycles of design research, and we do so according to the instructional sequence; Table 6.4 shows in which chronological order the activities have been designed, and which ideas presented in this chapter lead to the HLT tested in Chapter 7. In describing the design process, we show how activities developed by the Nashville team were revised and how ideas stemming from the historical and didactical phenomenology worked out in practice. Towards the end of the chapter, we also reflect on the use of the Minitools.

6.1 Outline of the hypothetical learning trajectory revisited

At the end of Chapter 5 we sketched an initial outline of an HLT that we summarized as challenging students to reason about aspects of distributions in increasingly sophisticated ways with increasingly sophisticated statistical notions and graphical representations. What we envisioned was that students would learn to reason coherently with the key notions of data, center, spread, sampling, and distribution. These notions then had to become more conventional and less dependent on context or graphical representations.

In the terminology of emergent models (2.1), we needed problem situations at a referential level that students could solve within a certain context with their own informal notions. In terms of statistical notions we thought of average in the daily sense, as well as informal notions of sampling, majority, spread, and consistency; all within the context students would be dealing with. From these activities at a referential level, *models of* the situations had to be developed, such as particular representations with which other problems could be solved (value-bar graph, for example). As pointed out in the Sections 4.3.1 and 5.3, we expected that estimation tasks would be useful for this level.

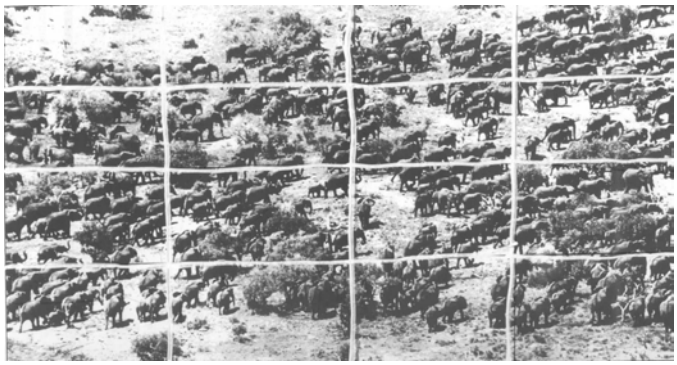
At a more general level, these *models of* had to become *models for* a more mathe-

matical reality, and students' informal notions had to become more conventional, supported by using Minitool 2. At some point we expected to address statistical issues as interesting in their own right and not necessarily bound to a particular statistical problem. For instance in their roles as data analysts, students could discuss the relative merits of graphical representations (6.5).

6.2 Estimation of a total number with an average

In the historical and didactical phenomenologies we argue the potential benefit of estimation tasks. As a starting point of the instructional unit, students had to invent a method of estimating the number of elephants in a picture (Figure 6.1). From the very start, we wanted to stress that we valued reasoning more than number answers or procedures. We aimed at challenging them to work with a notion of average in a way similar to the historical estimation examples (4.3.1).

a b c



d

Figure 6.1: Students' strategies to estimate the total number of elephants in the picture. (Reprinted with permission from Mathematics in Context © 1998 Encyclopaedia Britannica, Inc.)

The students in the seventh-grade classes used four main strategies with some variation (Figure 6.1). For each strategy, we add the frequencies in classes 1F (27 students) and 1E (28). The strategies were the following.

a Make groups, guess how many there are in each group, and add all numbers (15 + 10 + ...); this strategy was used by 0 students in 1F and 2 students in 1E.
b Make a group with a fixed number and estimate how many groups fit into the whole (in Figure 6.1b the students estimated groups of 10); this strategy was used by 6 and 11 students respectively.
c Count the number of elephants lengthwise and widthwise, and multiply these. Readers who have seen the video *Goodnight Mr. Bean* may recognize his method of counting sheep before falling asleep. We refer to this method as the 'area method' or the 'Mr. Bean method'; 4 and 2 students used this strategy respectively.
d Make a grid, choose an 'average box' and multiply this multiplicand by the number of boxes in the grid; this was used by 13 students in both classes.

Strategy *d* relies on an intuitive sense of average and, strikingly enough, this strategy led to numbers that were closest to the counted number of elephants (336). Strategies *a* and *b* yielded estimations that tended to be too low and strategy *c* tended to yield estimations that were too high. When we asked the students what they meant by an average box they described it as "a box with not too few and not too many." A similar description can be found in Aristotle's *Nichomachean Ethics* (Section 4.3.2). At this point we could not know how deep this description of an average box as "neither too few and nor too many elephants in it" was. We therefore looked for ways to let students become more explicit and precise: that is, ways to use their activity of estimating numbers with an average value as a model for more mathematical discussions. This issue is taken up in Section 6.6.
The estimation activity of elephants in a picture evoked the kind of reasoning that we had aimed for. In all four seventh-grade teaching experiments, students invented an 'average box' to estimate the total number of elephants. There are indications that students:
– indeed looked at the density of elephants or how the elephants were distributed in the picture at a referential level;
– seemed to like the activity (we inferred that from what they said about the activity and from how they worked at it);
– were later able to use this notion of an average box as a model for a more mathematical reality as interesting in its own right (Section 6.6).
For homework we asked the students to estimate the number of stars in a picture from the NASA website with the average box method, so that every student would get experience with this method (www.nasa.gov, picture of the day, April 29, 1999). From the experiences, we concluded that the elephant activity served as a suitable starting point for the HLT and that is how we used it in all subsequent seventh-grade teaching experiments. In short, students learned to use a notion of average to reduce

variation and were involved in reasoning about distribution and density at a referential level.

6.3 Estimation of a number from a total

In the didactical phenomenology we identified a variant of estimating a total number ($n \cdot \bar{x} = \Sigma$), namely estimating a number with an average value ($\Sigma / \bar{x} = n$). The activity we decided to use was the balloon activity (5.3):

> If normally eight adults are allowed in a certain balloon, how many seventh graders would you allow, if you only consider weight?

This activity would implicitly ask for an average that was not calculated but estimated. The estimated weights of these students and adults would function as a representative value and the activity could be used to let students think about sampling issues. In a class discussion these issues could then be made more explicit by asking the following questions: "How do you know the average weight of students? How can you make a reliable estimation of that?"

In all classes, students used two basic strategies to solve the problem. They all estimated the average weight of adults and seventh-grade students first. The first strategy was to use the proportion (e.g. 80 kg : 40 kg = 2) to estimate the number of students (2 * 8 adults = 16 students). The second strategy was to estimate the total weight that would be allowed (e.g. 75 kg * 8 adults = 600 kg) and divide that by the estimated weight of seventh-grade students (600 kg : 50 kg = 12 students). The answers in most classes varied from 10 to 16 students. Apparently, students knew the weight context well. When solving the balloon problem, some students in 1A asked a "normal looking child" how much he or she weighed; others asked a few students and took a value somewhere in the middle. One girl passed around a sheet of paper to collect others' weights, which we interpret as taking a small sample. Consequently, they were indeed dealing with the aspect of representativeness.

We concluded that the balloon activity could indeed support reasoning about average, representativeness, and sampling at a referential level. In Section 6.11 and 7.9 we describe how the balloon activity was used as a reference context for discussing sampling and distribution issues on a higher level. The balloon activity indeed functioned well in the beginning of the HLT to address the representativeness aspect of the mean and informal notions of sampling, which could be the basis for further discussion.

6.4 Talking through the data creation process

In all of the classes we used the battery problem in the second lesson to support students' reasoning about distribution aspects such as majority, outliers, and consistency (Section 2.3 and 5.1.2, and Figure 6.2). In the first trial (class 1A) we did not

spend much time on the introductory process of talking through the data creation process (P3 and R1). Soon we heard students say that they would choose the battery of 121 hours, which many of them also wrote down. We interpreted such answers as indications that students tended to focus on individual data values (2.2) and that the notion of sampling within this battery context was difficult for these students. We concluded that we had not spent enough time on talking through the data creation process. It was clear to us that we should spend more time on discussing the variable that has to be measured (life span), how it could be measured (with the same toy or walkman), and to look for alternative ways to deal with the sampling issue. Apparently, without a basic notion of measurement and sampling, students cannot really understand what the data values stand for. This supports R2 on the data creation process.

6.5 Data analyst role

One of the results of the Nashville experiments was the insight that students, in their role as data analysts, might reason on a more general level (R12). If students do not just solve a problem but also think about how to present the results to another person who could make a decision based on their result, they learn to reflect on the use of certain representations (at a model-for level). In this section, we present support for that claim and show how a seemingly subtle change in framing the question can lead to better reasoning. This is also an example of how the HLT changed between macro-cycles and how a result of the Nashville team was confirmed and extended in our research.

In the exploratory lessons in 1A, we asked the students in a straightforward way which brand was better, but they generally answered the question superficially. The students came up with one choice supported by one argument, for example "brand D because it has the highest value," "K is more reliable because it has four with the same value," or "brand K because it has more higher ones." Our instructional goal with presenting such problems to students was not that they were just to solve the problem, but that they develop techniques, notions, and graphs in order to learn to solve such problems in general and communicate about them. Students, however, cannot anticipate this goal when they read the simple question, "Which brand is better?" They probably think that they have to make a decision for themselves and choose between the two brands.

We expected students to give more profound answers and learn more if they took the role of data analysts (R12). Because many students in 1A found brand K "the best brand because it has four of the same values," we changed the four data values with exactly 115 into four slightly different values (113, 114, 115, and 116). If too many students were to choose one brand, we would miss the benefit of heated debates about which brand is best.

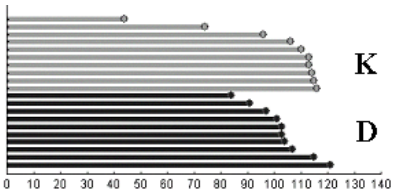


Figure 6.2: Battery life spans of two brands in hours (with slightly modified values)

In the first teaching experiment (1F), we framed the problem as follows.

> Write a letter to the shops on behalf of each factory. Give good arguments why the shops should sell your batteries.

Why did we choose for the letters from the factories? First, we wanted the students to find arguments for both brands. We thought that the conflict between several arguments could help the development of different measures of center and spread. Furthermore, we hoped that writing for a factory would make the students keener on the arguments, because there would be competition and money to earn.
Unfortunately, it did not work out this way. First of all, many students divided the task. One wrote the letter for brand D and the other student the other letter for brand K. In general, there was little interaction between the students. Second, some students used language that is common in advertisements: "You will see that D is just great" and "K batteries are the best!" They praised the batteries without many objective arguments and invented all kinds of additive information such as "not good, money back" (for brand K with the outliers). From a commonsense perspective, it is nice that students invent such solutions: if one brand happens to have outliers, giving the money back could be a solution. However, our hidden agenda with this activity was to let students reason about aspects of distributions in a more statistical way.
When the students worked on the problem, we heard a few of them refer to the *Consumer Reports* (*Consumentenbond*). One student even answered that the test was carried out by the *Consumer Reports*. This indicated that at least some students knew about the *Consumer Reports*, so for the second teaching experiment (1E) we decided to change the context.

> *Consumer Reports* has to report on the quality of these two brands. Give them such information that they can write about the two brands.

But we had additional reasons for choosing the *Consumer Reports* context. First, we expected that the students would discuss in pairs because they had to produce one report per pair. Second, we expected that the arguments would be more objective,

because the norm is that the *Consumer Reports* has to provide objective information. We expected the students to play the role of data analyst in a more serious or mathematical way than they did for the factories.
To compare the quality and quantity of students' arguments in the two conditions, we rated them (Table 6.1). The classes had the same level and preparation; the experiment in 1E was only six weeks later. The results in 1E were better than in 1F: in 1E there were many more arguments with a high rate and fewer arguments with a low rate than in 1F (Table 6.1). The argument that brand D was more constant or reliable only occurred in 1E. This result does of course not imply that students should always write reports to the *Consumer Reports*; the important thing is to look for a context in which students give objective and precise arguments. These arguments and ideas emerged while solving a specific problem; hence, we can characterize this activity as being on a referential level. The next step in the HLT would be to use the notions students developed during this activity in other situations, or in a more general or statistical way (cf. Doerr & English, 2003).

Table 6.1: Rated arguments about the two battery brands D and K

	low rate	**high rate**
	– D has the highest value – K has more values in a certain interval, e.g. 110-120 hours	– D has a higher mean than K – D: everything is close to the mean – D: everything above 80 – D is more consistent, reliable, or constant – K has outliers of a low value – K has more higher ones – Range of K is larger
1F	13 arguments with a low rate (by 24 students)	21 arguments with a high rate
1E	10 arguments with a low rate (by 23 students)	33 arguments with a high rate

With this example we demonstrated how one activity in the HLT changed in several design cycles: the roles of students switched from making a decision (choosing between two brands), to selling, and finally to being a data analyst. In the last role their arguments were most objective. At the level of instruction theory, the insights on the role of data analyst (R12) had received more empirical support and the importance of this role was substantiated by this example of changing the context.

6.6 Compensation strategy for the mean

Next we wanted to bring student experiences from the first two lessons together in the third lesson, by using the elephant context with the representation of Minitool 1

for estimating the mean visually. From the historical and didactical phenomenologies, we had concluded that a qualitative and visual way of working with averages is favorable to a procedural way of computing the mean (e.g. 5.4). Based on the exploratory interviews, we decided to challenge students to reinvent a compensation strategy for estimating means, as students had spontaneously done with a bar graph of the monthly temperatures (5.1.4). Because Minitool 1 resembles this bar graph representation, we expected that this representation with the value tool option would give students the opportunity to reinvent this method of compensating (see Figure 6.3). We used the activities of estimating elephants and the representation of the battery problem as reference contexts. We preferred the bar representation to balance models, because we assumed that the students would not have a good understanding of the physical laws of balance (cf. Hardiman et al., 1984).
In class 1F, the reinvention of the compensation strategy happened quite easily, but in 1E it did not work out very well, although the students in the two classes were considered to have equal learning abilities (as seen in their report grades for all subjects). We therefore analyzed the two lessons in more detail to find out what could have caused the difference.

Description of the third lesson in 1F
At the beginning of the lesson, the teacher and students discussed several strategies from the first lessons of estimating the number of elephants (Figure 6.1). One boy proposed to count the emptiest and the fullest box in order to find an 'average box', and multiply the number in these two boxes by 4 (there were eight boxes). The teacher remarked that this was the same as multiplying the 'average' by 8. This is a little more precise than "somewhere in the middle," which some other students said, but we aimed for a next step: using the average box for other, hypothetical situations.

Table 6.2: What would have been an average box in this case?

24	18	15
19	40	33
29	45	28
36	25	11

The teacher asked:

> Assume we had not estimated elephants but something else, what would have been an 'average box' here [in Table 6.2]?

This might sound strange since the numbers were already there, but this question was meant to let students explain what they meant by an average box. The students

did not have any trouble with this hypothetical question; the game-like activity drew their attention and evoked statistical reasoning. One boy said 29; he looked how many were above and below that number, which could indicate a median strategy. Others said (45 + 11) / 2 = 28, which shows a midrange strategy. A girl then used a counterexample to argue that the latter strategy was not reliable:

> But if you have one that is 100 and the rest [of the numbers] are 1, then you wouldn't take 50, would you?

From the reactions of her classmates we concluded that her point seemed clear. A boy then stressed that one has to look how the rest of the numbers lie in between the lowest and highest number. From such examples, we concluded that the activity of explaining what average boxes meant encouraged students to look at how the numbers were distributed–one of the things the HLT aimed at.
The next step in the evolving HLT was inspired by the didactical phenomenology: create situations in which the intuitive variants of center would create cognitive conflicts (cf. Watson, 2002) and ask for clearer definitions (5.3). The teacher showed another matrix with a more skewed distribution and asked the same question of what the average box would have been (Table 6.3). This time, students did not use the midrange anymore. Instead, some students looked for a bunch of numbers that were close together, which is similar to what Noss et al. (1999) report about nurses interpreting graphs on blood pressure. Others looked at how many were above or below a certain number, which could be seen as a precursor to the median. In short, the discussion on what an average box might mean led to the reinvention of measures of center such as midrange and median, though the students did not have words for them yet.

Table 6.3: A more skewed distribution of numbers. The midrange was rejected.

35	58	91
93	83	89
98	97	68
76	82	11

After this activity of explaining what students meant by an 'average box' with the matrices, we returned to the elephant problem with the same question, but with a different representation to evoke a compensation strategy. The teacher showed a value-bar graph of the numbers of elephants in the boxes (Figure 6.3).

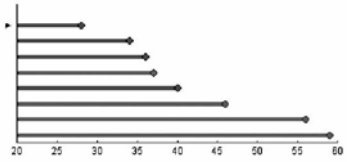


Figure 6.3: What was the average box if you look at this Minitool 1 representation?

The students discussed this question extensively. First they talked about the number of values before and beyond 40 or 45. In order to push the discussion into the direction of compensating, the teacher drew a vertical line at 40 (Figure 6.3). This is an example of her proactive role while guiding the reinvention process (cf. McClain, 2000). Then the students started to reason about the lengths of the bars. One girl drew a circle around what was too high on one side and too short on the other. "What you cut off on one side you give to the other side." This convinced the other students that 40 was too little. "But 45 is too much," others said. One boy proposed a value somewhere in between, 42, which was in fact the real average, 336 / 8 = 42. It struck us how precisely students could estimate the mean with value-bar graphs.

This is how the strategy of compensation arose in 1F without much effort from the teacher's side. The strategy was easily adopted and applied in other situations such as the battery problem. A few students even employed this strategy as their favorite for comparing data sets. This implies that the grid with average boxes first functioned as a model of the elephant problem and later as a model for comparing data sets. In this way, students reinvented informal measures of center including the midrange, a precursor to the median, and a value somewhere in a cluster. (This last strategy was common among Greek astronomers as well.)

From the fact that a student chose 42 as the mean, it is clear that these students understood that the mean need not be one of the values, in contrast to what students of this age often think (Mokros & Russell, 1995; Strauss & Bichler, 1988). We assume that this was stimulated by the visual, continuous way of dealing with the mean in Minitool 1. This way of dealing with the mean avoids the algorithm and seems to be more connected to qualitative aspects such as intermediacy, compensation, and representativeness. Therefore, we concluded that this strategy was an important part of the HLT and we wanted to replicate this success in 1E.[20]

Unfortunately, evoking the compensation strategy did not happen as readily in 1E as in 1F. The cause was not that the students of 1F were smarter: other things went bet-

20. Students in Nashville did not use the value tool for estimating means (R9), except one student. In that case the estimation strategy was a way of calculating, whereas in the case that was just described, the goal was not to calculate the mean but to explain what the average box would have been.

ter in 1E (see Sections 6.5 and 6.11). Rather, it turned out that we had not analyzed our initial success in 1F well enough.
After analyzing both lessons we concluded that the teacher had not always asked the right questions and the HLT had not provided her with suitable questions. Where the main question in 1F was to explain what students meant by an average box, the teacher asked in 1E:

Teacher: How could you estimate the mean without calculation?
Kristin: Take the number that is closest.
Bas: It is around the [*inaudible*].
Teacher: How could you check this?

Finally, after many attempts by the teacher to guide the discussion, one student mentioned the leveling-out strategy that we had hoped for.

Anissa: The long ones that stick out at the right can fill up the other parts.

The teacher then explained how this worked.
In 1F the goal was to explain what students meant by an 'average box' in relation to the elephant problem while using the representation of the battery problem. The method of compensation arose as a way of explaining and justifying what an average box was. In 1E, however, it was not clear what the reference context was, the elephants or the battery problem.
This implies that both the designer and teacher should be very aware of where they want to go, and more importantly where the students come from, so that the designer and teacher are also aware of the previous knowledge upon which they can build. In other words, the HLT should be very explicit and clear for the teacher. In this example, our HLT was not explicit enough. Somehow, while developing the HLT we tended to focus on things that did *not* go as we anticipated. Due to this incident in 1F and 1E we came to acknowledge the importance of learning from design success as well. In the remaining teaching experiments we would try to repeat what had happened in 1F by asking what would have been an average box in hypothetical cases.

6.7 Data invention in the battery context

One of the prime goals of the HLT was that students would come to see a data set as a whole instead of just as a collection of individual data points. During the battery activity students had already reasoned about how good and reliable the brands were, but these informal notions did not seem to match with average and spread. As a way of assessing their statistical understanding and to stimulate the average coming to stand for 'how the batteries last', and spread for reliability, we decided to reverse the battery problem in 1C. Around the fourth lesson, we asked the students to invent data sets in Minitool 1 that belonged to the particular characteristics of hypothetical brands. We asked the students to invent data sets in Minitool 1 that belonged to the

following propositions:

– Brand A is good and reliable.
– Brand B is good but unreliable.
– Brand C's spread is like that of A but the brand it less good.

We conjectured that going back and forth between the Minitool 1 representation and notions such as reliability and spread would foster a stronger link between the symbol system of value-bar graphs and the conceptual network of statistical notions (cf. Sfard, 2000b; Steinbring 1997).

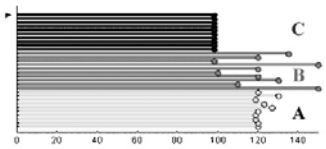


Figure 6.4: Invented data for brands A, B, and C (ten values per brand)

We did not specify a sample size, but most students chose ten per brand (as in the first battery problem). Almost all of the students made something like Figure 6.4, which roughly shows a match between average and 'good' and between spread and 'reliability' (in this example, the spread of brand C is not similar to that of A). In other words, the students gradually came to use more statistical notions. From the notes we made in the classroom and from the videotapes, we observed that most of the students liked this activity and that they were actively involved.

We concluded from the retrospective analysis that the back-and-forth movement between networks of notions and graphs, and presumably between graphs, would be a good design heuristic for several reasons.

1 Lemke (2003) writes that students can express particular ideas on quantitative phenomena more easily with graphs than with words. Consequently, if students express their ideas of quantitative phenomena in graphs, it might be easier for teachers and researchers to assess what students really understand. This could be more than they express in words.
2 We can use students' cognitive limitations to prevent them from thinking of a bunch of individual data values, namely by asking for aggregate characteristics. Konold and Higgins (2003, p. 203) write, "With the individuals as the foci, it is difficult to see the forest for the trees." Inspired by this metaphor we formulated the following design heuristic for statistics education:

   > If students do not see the forest for the trees, ask about characteristics of the forest or other forests.

   Here we asked for a characteristic "good but unreliable" as an overall character-

istic of a data set that was to be constructed. In this way we created a need for a cognitive structure that could help to see a data set as a whole or a composite unit.

3 In most school textbooks, students have to interpret ready-made graphs (Friel et al., 2001) and if they have to make their own graphs, the type of graph and drawing procedure are often prescribed. De Lange et al. (1993) and Meira (1995) recommend letting students invent their own graphs and inscriptions. We assume that this freedom to make their own symbolizations will enlarge the chance that these are meaningful and functional to students.
In retrospect, we were positive enough about the results of this reverse battery activity to use the idea again in 1B.

6.8 Towards sampling: Trial of the Pyx

In the historical and didactical phenomenologies, we conjecture that a context such as the Trial of the Pyx (4.4.3) would be useful to let students reinvent sampling methods (H10). In the case of estimation we translated the historical contexts into modern ones: instead of the number of leaves and fruit on a tree we asked to estimate the number of elephants in a herd. In 1A and 1F, however, we tried to use a historical example without translation into a modern context. The text we gave the students in 1F was:

> A long time ago, coins were made of gold and silver. The coins had the same value as the value of the gold or silver, which is different from the present situation. Nowadays they use cheaper materials to make coins. From the 12th to the 18th century coin makers had to be checked: they could have used too much or too little gold or silver for the coins. If they used too little gold they were fired and had to pay a fine to the King. If they used too much gold they spoilt the King's gold and were punished too. It was not possible to weigh every coin because that was too much work.
>
> Describe extensively your advice to the King. How should he let the coins be checked? Explain why your method is good.

With a summary of an episode from 1F, we intend to give an impression of students' ideas of sampling in such a context and to give an example of an activity that we did not keep in the HLT.
The teacher asked the students what they would advise the King to check the coin makers. At first she did not get a clear answer, so she asked what the problem was.

> Teacher: Can you explain what the problem was with the coins?
> Marek: Yes, that they didn't weigh the same, these coins.

This means that Marek realized that there was variation in the coins' weights; we consider this the basis of statistical investigation (cf. 2.2). Joan then made a connection with the mean.

Joan: You take 25 coins, weigh them and divide by 25. You do the same for all and then you know how much one coin weighs, so you don't need to weigh every coin separately.
Teacher: Why do you take 25?
Joan: You could also take 50.
Teacher: OK, but why do you do all this?
Joan: Then you can calculate the mean.
Teacher: When you have found the mean, what then?

Indeed, Joan's idea is time-saving compared to weighing individual coins and it demonstrates understanding of the mean as an aggregate feature of a group of coins. The intention of the HLT, however, was that students would find out how these coins should be selected, because that would nicely leap into the issue of random sampling. Unfortunately, the teacher did not ask this particular question of how the coins should be selected (the HLT had not been explicit about this), but focused on the question of how the coin makers had to be checked. In this class, students clung to computations of means, though we tried hard to direct their attention towards more qualitative aspects of the average values. Students proposed and repeated different strategies but they wanted to test all coins. None of them went in the direction of sampling until two boys had the following ideas.

Timo: Well, you can burst into the coin makers' place, take a balance, take 100 coins or so, and weigh them. If the result is not 10 grams per coin, then it is wrong. Then you have to punish them.
Jelle: A spy, put a spy into the smith's place. Give him a piece of 20 grams gold that is surely 20 grams.

Although we thought that the discussion was useful preparation to developing a notion of sampling, we concluded that the historical context formed an extra problem to overcome (cf. Van Amerom, 2002). Most students did not know what the situation was in England around 1200; instead of being a statistical problem this also became a historical one. Students from class 1A even talked anachronistically of computers and ingenious machines that would throw out inferior coins. We also concluded that these students had no clear intuition about sampling in such a context. Based on these two conclusions, we decided not to use this context again, but to look for other ways of letting students think about sampling.

6.9 Median and outliers

The Nashville team capitalized on the median as a measure of center, but it turned out to be difficult for students to develop an understanding of the median as a measure of center of distribution (R16). As motivated in the historical and didactical phenomenologies, we capitalized on the mean, but still thought that the median should somehow be addressed. One trivial reason is that the median is in the Standards (NCTM, 2000) and in the Dutch mathematics curriculum of grade 8 (Meth-

odewijzer, 1998). The main reason was that the median is a useful characteristic of skewed distributions and data sets with outliers; it is less sensitive to outliers (more robust) than the mean, which is particularly helpful in social or economic contexts with irregular data (Chapter 4). We expected that if students dealt with the difference between mean and median that they would also think about aspects of distribution such as outliers and skewness.

The first activity with which we tried to challenge students to use the median was the so-called wing span activity (in Minitool 1). Students had to inform a bird encyclopaedia about the size of adult birds (sparrow, blackbird, sea eagle, albatross) on the basis of data sets, two of which included suspect values (a sparrow with 0.15 cm wing span and two young albatrosses). The intention was that students would see that the middle-most value (median) is useful if there are outliers, but we also would appreciate ranges in the middle of the data set.

It did not quite work out that way. Although students came to think about outliers, they mostly did not realize the usefulness of the median. Some kept using the mean by estimating it with the value tool. A few students just took a middle range, which is very sensible in this context because it accounts for the variation that occurs in real life: "Adult albatrosses have a wing span of 330-360 cm." The median was defined as the middle-most value with respect to value-bar graphs and dot plots. However, from later mini-interviews, we concluded that the vast majority of students confused the median with the midrange. And if they developed some understanding of the median in relation to the two equal groups of Minitool 2, some thought it as always different from the mean. From the final interviews in class 1F we concluded that even those students who were able to find the median in Minitool 1 were not able to find it in a series of numbers. Apparently their understanding of the median depended heavily on the representation and context they were used to. We decided to try and combine the two equal groups option with series of numbers in discussions about mean and median (Section 7.7).

In short, we encountered many problems with the median and could not think of a suitable problem situation that really begged to be organized by the median. Yet we considered the median an important measure of center to address because, by addressing it, we could make the distribution of data values a topic of discussion. We decided to have a closer look at the history of the median to find indications for the conceptual problems and to find phenomena that could inspire us for useful problem situations (4.5).

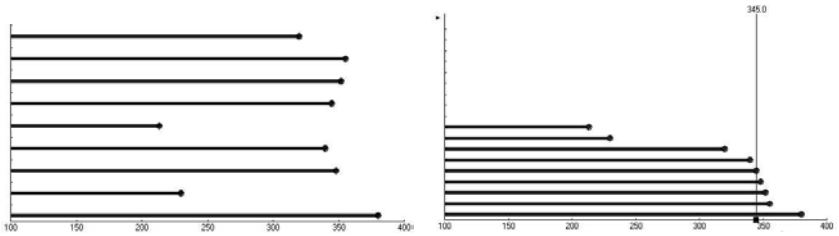


Figure 6.5: Wing spans in cm of nine albatrosses in Minitool 1. Left is unordered and right is ordered with a value tool at the median, 345 cm; in both figures, another subset of data is hidden by the 'hide' option

6.10 Low, average, and high values

When designing an HLT it is important to ascertain which prior knowledge students already have and which knowledge they can easily develop. In the beginning of the teaching experiments in grade 7, we did not know what intuitions students would have of distributions. It was not until we interviewed one student in 1E that we were able to formulate a conjecture about this that was confirmed in all subsequent teaching experiments.

During the ninth lesson in 1E after discussing the balloon activity, students had to make a graph of students' height or weight for the balloon driver. In a mini-interview, we asked Danny what his prediction would be before collecting any data. He started to draw the first sketch in Figure 6.6. We initially thought that he was drawing a hill shape. When we asked for clarification, he did not answer but drew the next sketch and then said, "There are taller and shorter and average students," while pointing at the corresponding parts in the sketch. We wondered if he saw the students standing next to each other, so we asked him where he got this idea from. He then made a third sketch, which resembled the dot plot of Minitool 2. Again, he explained that there were short, average, and tall students while indicating the various parts of the dot plot. From his sketch we can clearly see how he conceptually organized the height phenomenon with three separate groups. He said that there were "more around average," by which he presumably meant a high density or frequency in a certain interval. This led us to conjecture that students easily come to organize data sets into three groups of low, average and high values if they know the context (in which a normal distribution is to be expected).

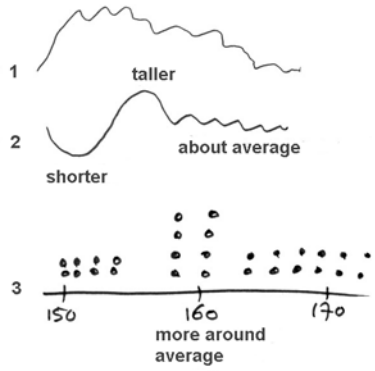


Figure 6.6: Danny's organization of the height distribution into three groups

This mini-interview also provides an example of symbolizing, which we characterized for the time being as the process of making a symbol for a specific purpose, using it, improving it and possibly making a new symbol (2.3.4). To clarify his thoughts, Danny made a sketch, which he improved and explained, and again changed into another symbol with a more explicit explanation (cf. Lehrer, Schauble, Carpenter, & Penner, 2000; Meira, 1995). Again we see the importance of students' symbolizations and their explanations: when analyzing their inscriptions and reflections we can better understand how they think (cf. 6.7). In Section 8.2 we analyze this mini-interview as an instance of diagrammatic reasoning.

6.11 Reasoning about shape

6.11.1 Anticipations in the HLT

One of main goals in the HLT was that students would learn to reason about the shape of distributions (Section 2.2 and R14). Though students in the first two teaching experiments started to reason about majorities and density when using Minitool 2, for example during the speed trap activity, they did not explicitly reason with shape. We had hoped that they would reason with 'hills', as reported by the Nashville team (R10), but they did not. A possible reason for this is that the reasoning with hills in their teaching experiment occurred in the final phase of 34 lessons, whereas our teaching experiments lasted only 12 or 15 lessons. In this section, we show how students in class 1E came to reason about and with shapes.

Because De Lange et al. (1993) and Meira (1995) report positively on student invented graphs, and because of the RME tenet that promotes student productions and

constructions (Treffers, 1987), we decided to let students invent their own graphs for their own data (cf. 6.10). We expected that their own graphs would be more meaningful and functional for them than ready-made representations (cf. 6.7). In addition, we thought it would help if students worked with their own data in a familiar context. To avoid a focus on individual data, we also played with the idea of activities using sketch-like graphs without any data. As a follow-up to the balloon activity (Section 6.3), we asked the students to make a graph for the balloon driver with which she could decide how many students she could safely take on board. Note that students were encouraged to take a data analyst role in this question (6.5).

6.11.2 Retrospective analysis

The students drew various graphs (Figure 6.7), but the teacher focused the discussion on two of them, namely Mike and Emily's graphs (Figure 6.8 and 6.9), presumably because she saw an opportunity to talk about hills with the help of Mike's graph.

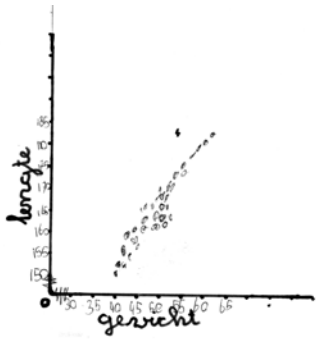


Figure 6.7: A scatterplot that was not discussed

She postponed Bas's graph, which resembles Minitool 3, but did not manage to discuss it. (We could interpret Bas's creation as a reinvention of the scatterplot, which the students had not learned in mathematics lessons.)

Figure 6.8: Mike's graph

Mike explained how he got the dots as follows.

Mike: Look, this is roughly, averagely speaking, the amount of students with this weight and there I have put a dot. And then I have the number of students on the left [*y*-axis]. There is one student who weighs about 35 [kg], and there is one who weighs 36, and two who weigh 38 roughly.

And so forth: the dot at 48, for example, signifies about four students with weights of around 48 kg. After some other graphs had been discussed, including that of Emily, the teacher asked the following.

Teacher: What can you easily see in this graph [Mike made]?
Laura: Well, that the average, that most students in the class, uhm, well, are between 39 and, well, 48.
Teacher: Yes, here you can see at once which weight most students in this class roughly have, what is about the biggest group. Just because you see this bump here. We lost the bump in Emily's graph.

Apparently, Mike's graph helped students see the majority of the data–between 39 and 48 kg. This 'average' or group of 'most students' is an instance of what Konold and colleagues (2002) call a 'modal clump'. Teachers and curriculum designers can use students' informal reasoning with clumps as preparation to their using the average as a representative value for the whole group, for example.
Here, the teacher used the term 'bump' to draw students' attention to the shape of the data. By saying that "we lost the bump in Emily's graph," she invited the students to think about an explanation for this observation. Nathalie reacted as follows.

Nathalie: The difference between … they stand from short to tall, so the bump, that is where the things, where the bars [from Emily's graph] are closest to one another.

Teacher: What do you mean, where the bars are closest?
Nathalie: The difference, the endpoints [of the bars], do not differ so much with the next one.

Evelien added to Nathalie's remarks:

Evelien: If you look well, then you see that almost in the middle, there it is straight almost and uh, yeah that [teacher points at the horizontal part in Emily's graph].
Teacher: And that is what you [Nathalie] also said, uh, they are close together and here they are bunched up, as far as (…) weight is concerned.
Evelien: And that is also that bump.

These episodes demonstrate that, for the students, the bump was not merely a visual characteristic of a certain graph, but that it signified a relatively large number of data points with about the same value–both in a hill-type graph and in a value-bar graph. For the students, the term 'bump' signified a range where there was a relatively high density of data points, which they referred to as the 'majority'.

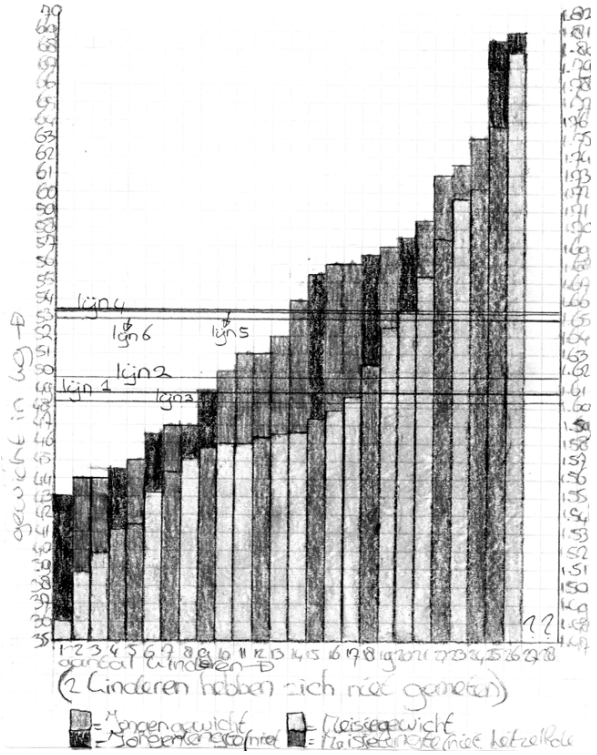


Figure 6.9: Emily's graph. The lighter, smaller bars represent students' weights; the darker bars students' heights. (Although all students used the same data set, Mike's graph does not exactly match the values in Emily's graph. Mike's graph is more like a rough sketch.)

In the next lesson, students used the bump as a reasoning tool, as the next episode shows.

Laura:	But then you see the bump here, let's say [Figure 6.10].
Yvonne:	This is the bump [pointing at the straight vertical part of the lower ten bars].
Researcher:	Where is that bump? Is it where you put that red line [the vertical bar]?
Laura:	Yes, we used that value bar for it (...) to indicate it, indicate the bump. If you look at green [the upper ten], then you see that it lies further, the bump. So we think that green is better, because the bump is further.

The examples show that some students started to reason about shape, which was indeed the purpose of these activities. However, they still focused on the majority, the modal clump, instead of the whole distribution. This seemed to change in the thirteenth lesson.

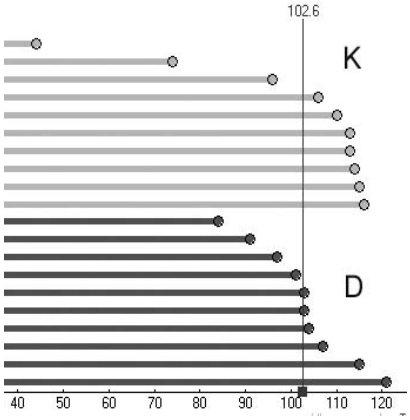


Figure 6.10: Reasoning with the 'bump' in Minitool 1

In that lesson, we discovered that asking students to predict and reason without available data can be helpful in fostering an aggregate view of data. An initial example of such a question was to predict what a graph of the weights of eighth graders might look like, as opposed to one of seventh graders. We hoped that students would shift the whole shape instead of just the individual dots or the majority.

Teacher: What would a graph of the weights of eighth graders look like?
Luke: I think about the same, but another size, other numbers.
Gerdien: The bump would be more to the right.
Teacher: What would it mean for the box plots?
Mike: Also moves to the right. That bump in the middle is in fact just the box plot, which moves more to the right.

It could well be that Luke reasoned with individual numbers, but he thought that the shape would remain the same. Instead of talking about individual data points, Gerdien talked about a bump, in singular, shifted to the right. Mike related to the box plot as well, though he probably just referred to the box of the box plot.
Another prediction question also led to reasoning about the whole shape, this time in relation to other statistical notions such as outliers and sample size. Note that students in all classes used the term 'outliers' for very low and high values, whereas statisticians only use the term outliers for exceptional or suspect values that fall outside the distribution.

Researcher: If you measured all seventh graders in the city instead of just your class, how would the graph change, or wouldn't it change?
Emily: Then there would come a little more to the left and a little more to the right. Then the bump would become a little wider, I think. [She explained this using the term 'outliers'.]
Researcher: Is there anybody who does not agree?

Mike: Yeah, if there are more children, then the average, so the most, that also becomes more. So the bump just stays the same.
Anissa: I think that the number of children increases and that the bump stays the same.

In this episode, Emily related shape to 'outliers'; she thought that the bump would grow wider if the sample grew. Mike argued that the group in the middle also grew higher, which for him implied that the bump kept the same shape. Anissa's answer is interesting in that she seemed to think of relative frequency: for her the shape of the distribution seemed to be independent of the sample size. If she had thought of absolute frequency she probably would have thought that the bump would be much higher. Apparently, the notion of a bump helped these students to reason about the shape of the distribution in hypothetical situations. In this way, they overcame the problem of seeing only individual data points.
A last example illustrates how two students came to reason about distribution. As with Laura and Yvonne (Figure 6.10), they were not disturbed by the fact that distributions do not literally look like hills in Minitool 1. In the final test, students had to revisit the battery problem with more advanced questions such as whether the distributions of the battery brands looked 'normal' or skewed. 'Normal' was informally defined as "symmetrical, with the median in the middle and the majority close to the median." They used the term 'hill' to indicate the majority in Minitool 1 (see Figure 6.10).

Anissa: Oh, that one [brand of lower ten bars] is normal (…).
Nathalie: That hill.
Anissa: And skewed if like here [the upper ten bars] the hill is here [the straight part].

Again, this indicates that the notion of a hill was not just visual but had become a conceptual tool for different students.
From these lessons in 1E we were able to draw a number of conclusions. Our first conclusion was that it can be useful to let students invent their own graphs of their own data to allow a discussion on the merits of the different graphs.
Second, we observed that reasoning of high quality only occurred during the lessons without computers. This raised the question of what the role of computer tools is for students' learning. Do the tools constrain their thinking or are students not inclined to reflect when working with a computer unless explicitly asked to? We conjectured that the students' experience with the Minitools was very influential; their graphs (except Bas's) were clearly influenced by the Minitools background. From the mini-interviews in this class, however, we concluded that on the whole the students themselves did not see the resemblance with the Minitools. We also conjectured that reflective discussion is easier without computers.
Third, we concluded that predictions about shape in hypothetical situations can be

helpful to foster understanding of shape or distribution. If students predict a graph without having data, they have to reason more globally with an aggregate feature in mind (earlier we defined the design heuristic, "Ask questions about the forest, or predict properties of other forests"). In this way, designers can use humans' cognitive limitations: if there are no available data and students have to predict something on the basis of a conceptual characteristic, it is impossible to imagine many individual data points. Another, more slogan-like heuristic that we used was, "sometimes stay away from data."

The type of reasoning developed in this class 1E came closest to what we aimed for in the HLT. To learn from the success (cf. 6.6) we decided to analyze students' reasoning about bumps more extensively and answer the second research question on the basis of the episodes of this section (see Chapter 8).

In the HLT we had not explicitly dealt with the relation of sample to population; we thought that this distinction would be too technical for students at this stage. In hindsight, we acknowledged that the idea of reasoning about very large samples comes close to linking the concepts of sample and population. For the eighth-grade teaching experiment we decided to make this distinction explicit (Chapter 9).

During the teaching experiments we often wondered whether the goal of distribution as an object-like entity would not be too demanding for the majority of students. Sometimes students themselves hinted at that. For instance, two students asked us, "Why do we have to solve those adult questions?" And indeed, when we visited a class during another lesson it struck us that the teacher treated them much more as young children than we did (reading aloud a story by Roald Dahl, for instance). We took this as one of the indications that the end goal was too demanding for the majority of the students. We therefore decided to focus on simpler issues such as spread for the teaching experiment in 1B and to do the next macro-cycle on distribution with older students (grade 8).

6.12 Revision of the Minitools

Soon after the first lessons with the Minitools, we formulated a list of revision wishes. The most important were:

- Change the slow applets, running via the Internet, into faster stand-alone applications that also provide options to save and print.
- Make different versions of Minitool 1 and 2, with fewer and more options. The main reason for this was that we wanted students to focus on the global shape (cf. Ben-Zvi & Arcavi, 2001) and not make vertical slices with fixed interval width for instance. We had seen examples of this in our own experiment as well, for instance with the speed trap.
- Make histogram and box plot available after students have chosen equal interval width or four equal groups respectively. We found it unsatisfactory that students

were prepared to use histograms and box plots, but would never be able to use them within the software. If we want to support students to go from organizing data in certain ways to conventional representations, then we should incorporate these graph options (see Figures 2.5 and 2.7).

Unfortunately, it took until the last seventh-grade experiment before these revisions were realized. It also took until after that experiment before the revised Minitools worked well at the school site. Nonetheless, but we had the revised Minitools at our disposal for the teaching experiments in grade 8.

6.13 Is Minitool 1 necessary?

In reaction to a presentation of the Minitools, a few colleagues questioned the function of Minitool 1 and some found it more difficult to interpret than Minitool 2. From the exploratory interviews (5.1) and from the first two teaching experiments, we had no clear evidence whether Minitool 1 or 2 would be easier to use for students. To find out whether using Minitool 1 had any added value, we decided to start with Minitool 2 in one *havo*-class (1C) and compare what happened with the next *havo*-class (1B), hoping that those classes would be comparable. Classes 1C and 1B took the same pretest and participated in the same activities in the first couple of lessons, except that 1C solved the first battery problem with Minitool 2, and 1B with Minitool 1. Unfortunately, class 1C turned out incomparable with the next *havo*-class: the students did not concentrate well and had lower results; this also held true for subjects other than mathematics. Hence, we could not draw reliable conclusions by comparing the classes. Yet there were some indications that Minitool 1 has its use. We give a few examples in the remainder of this section.

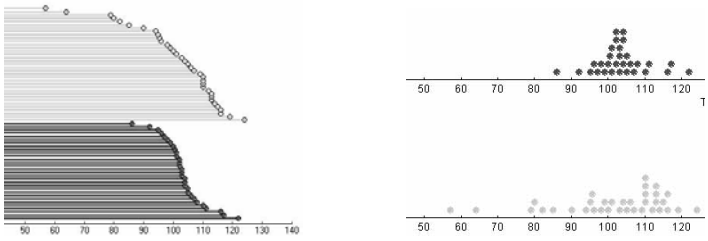


Figure 6.11: Battery problem in Minitools 1 and 2, now with a larger data set

After the students in 1C had worked with Minitool 2 in the second lesson, the teacher introduced Minitool 1 in the third lesson and asked about the average. One student spontaneously exclaimed, “But those bars are much easier!”

In the fourth lesson, when we interviewed him about the merits of the two Minitools, he said that he found Minitool 1 handier [*handiger*] and Minitool 2 more obscure

[*ondoorzichtiger*] but could not explain why. In the eighth lesson, we interviewed other students, working on the speed trap problem. Some of them found Minitool 1 better organized [*overzichtelijker*], whereas others preferred Minitool 2, and it turned out that there could be a difference in what the tool was used for. Consider this example.

Carl:	I mean, such a dot, you can hardly see it.
Mark:	Indeed. (...)
Interv.:	Can you explain in more detail why you find this handier than Minitool 2?
Carl:	It is harder to calculate the mean [in Minitool 2 than in Minitool 1]. (...)
Mark:	Yes.
Interv.:	And what could you see better here [in Minitool 2], perhaps?
Carl:	Spread.
Mark:	Yes.
Carl:	Yes, you see that with the other one [Minitool 2] (...) You see the spread because it is so [moves his finger from the lowest to the highest value].
Mark:	From here to here [from lowest to highest value].

Not all of the students found Minitool 1 better organized than Minitool 2. Consider the following excerpt:

Interv.:	What did you find clearer, this graph with the dots or that graph with the bars?
Martin:	The other one [Minitool 2 with the dots]. That is, I found that one better because you can see better where everything is. Because here [Minitool 1] you do not get a really clear image of it.

However, like the earlier pair, this pair found Minitool 1 better for estimating means:

Interv.:	Where was it easier to estimate the mean?
Martin:	The mean was easier to estimate in that one [Minitool 1].
Interv.:	And why is that?
Martin:	You can stick them together [*kan je het op elkaar plakken*].
Edward:	Yes.
Martin:	You can turn them and stick them to each other.
Interv.:	And why can't you do that with the dots?
Martin:	Because there are no lines.
Edward:	Yes, they are all dots.
Martin:	It is a different graph.
Interv.:	And why is the mean so difficult to estimate here [in Minitool 2]?
Edward:	You cannot stick it on the left-hand side [*omklappen*].
Martin:	You cannot do [in Minitool 2] what we do here [sticking together in Minitool 1] as easily.
Interv.:	And where do you better see how the spread is?
Martin:	With the other [Minitool 2].

From such episodes, we conjectured that students would find Minitool 1 easier for

estimating means and Minitool 2 for looking at how the data are spread out. Because there were only a few mini-interviews supporting this conjecture, we decided to test it in class 1B. The other interviews in 1C give a mixed impression without an unequivocal conclusion. One of the reasons is that many students seemed to interpret our question as, "In which Minitool is it easier to read off the values?"
The issues raised in this section lead to the questions of what the affordances of both tools are, and what the value of comparing two representations is (for theories on multiple representations see Ainsworth, Bibby, & Wood, 2002; Van Someren, Reimann, Bozhimen, & De Jong, 1998). In that sense, Minitool 1 could well be useful in addition to Minitool 2. As the interviews tentatively indicate, both Minitools may support particular types of reasoning in different problem situations. To gain more insight into this issue, we decided to ask students more about the two Minitools in the last seventh-grade teaching experiment.

6.14 Reflection on the results

In this section we reflect on the results of the various teaching experiments analyzed here (1F, 1E, and 1C) and compare them with those of the Nashville research. This makes it easier to formulate what in the evolving instruction theory should be rejected, refined, or has been confirmed. The results presented below reflect the patchwork character of the design process. At the end of this section we reflect on the first outline of the HLT (5.4) and indicate which activities were to be used in the last teaching experiment in grade 7 (Chapter 7).

1 *Intertwinement of key concepts*. Our point of departure as stated in Section 2.4 and the didactical phenomenology (Chapter 5) is to deal with several key concepts of statistics at the same time, starting at a referential level (2.1). The activities we used in the beginning of all teaching experiments, elephant estimation, balloon, battery problem, all served that purpose. These activities formed the basis for students to reason about representativeness, skewness, the distribution of number (or elephants), majority, 'spread out', reliability, and basic sampling in specific contexts. When working with Minitool 2, students dealt with additional notions that were discussed such as frequency, density, median, range, and spread. In 1E students came to reason about shape in relation to statistical notions such as outliers and sample size.
2 *Mean*. In the historical and didactical phenomenology, we argue that the mean should play an important role in an instructional sequence for early statistics education. As apparent from the different teaching experiments, students are able to reinvent different measures of center if this process is well guided. Though we understand why the Nashville team wanted to avoid the mean (P7), we still advocate attention for the mean from an early stage onwards. We must however account for the cultural difference: Dutch students are apparently not as drilled to

use the mean for statistical problems as American students and have a better understanding of the mean (5.1).

3 *Compensation strategy*. We show how students reinvented a compensation strategy of finding the mean with case-value bars. We are convinced that this representation is more suitable than the common balance metaphor. As Hardiman et al. (1984) note, this balance metaphor is only useful if students know that physical context from science classes or daily life (see also Pine & Messer, 2000). These seventh-grade students had not yet learned about this at school, though they probably had informal knowledge about seesaws.

4 *Median*. Despite some examples of students reinventing the median method for comparing data sets, we had not been very successful in designing good instructional activities for developing a notion of the median as a measure of center. The vast majority of students confused it with the midrange. Students in 1F had developed some understanding of the median in relation to the two equal groups of Minitool 2, but they did not see it as a measure of center, they tended to see it as always different from the mean, and they could not find it in a row of numbers. We decided to combine the two equal groups option with series of numbers in discussions about mean and median (6.9) and to study the history of the median.

5 *Talking through the process of data creation*. This is indeed important as we rediscovered, which supports R1, and it is indeed necessary as a way to address sampling issues (R2). We assume this result, like some of the other results of this section, can be taken as part of an instruction theory for statistics education.

6 *Role as data analyst*. The results of the battery activity in 1A, 1F, and 1E show that students were more objective and precise in the *Consumer Reports* than in the factories context. This supports and substantiates R12 (about the importance of the role of data analyst), which we take as part of our the instruction theory as well. We are well aware that it is still necessary to establish a socio-mathematical norm (Yackel & Cobb, 1996) of taking that role (2.3). Letting students produce a report to the *Consumer Reports* is not enough to establish such a norm.

7 *Data invention.* To establish a closer relationship between notions such as average, quality, reliability, and spread on the one hand, and graphs on the other, we asked students to invent their own data sets that would fit certain characteristics, in this case aggregate features of battery brands. We saw this as a successful instance of a back-and-forth movement between notions and graphs, inspired by theories on symbolizing. The key issue is that students can develop conceptual structures when inventing data sets that have particular aggregate features.

8 *Shape of the distribution*. It was not easy to evoke reasoning about shapes of distributions, but in 1E we succeeded by letting students make their own graphs of their own data and discuss a few student graphs. In this reasoning about shape, students used several statistical notions (outliers, majority, average) to explain

how the shape would change in hypothetical situations in a well-known context. Clearly, the 'bump' had become more than just a visual feature and served well in reasoning about aggregate features of data sets (as opposed to a case-oriented view on data). To understand students' development of a shape notion in a more general way, we decided to analyze these episodes on reasoning with the bump more extensively with semiotic theories that are not domain-specific.

9 *Low, average, and high values.* We have seen many examples in which students conceptually divide data sets of unimodal distributions into three groups of low, average, and high values. This categorization into three groups is already better than looking at individual data values. We assumed that students already have an intuition about distribution in that three-groups sense and that the activities of this sequence helped them to express this intuition with notions and graphs.

10 *Histogram and box plot.*[21] A mistake that some students made with histogram-type graphs was that they interpreted the height of bars as signifying height of people instead of the frequency of people with that specific height. Many students found box plots hard to use and understand, though some could describe advantages of both graphs. For the teaching experiment in 1B we decided to focus on spread and sampling, and not so much on conventional graphs such as the histogram and the box plot.

11 *Design heuristics*. In this chapter we formulate several design heuristics that are partially related to each other. If students do not see the forest for the trees, ask them about the forest or other forests; go back and forth between graphs and notions, or different graphs; predict the shape of distributions in hypothetical situations; and sometimes stay away from data. These heuristics were applied again in subsequent teaching experiments.

12 *Is Minitool 1 necessary?* There were indications that using Minitool 1 in addition to Minitool 2 made a difference. One successful change in the HLT was to let students compare the same data set in both Minitools when we introduced Minitool 2. We also offered students the freedom to solve a problem in the Minitool they preferred to use. In our view, the original chain of signification was too linear for our purpose of designing and revising the HLT. More on this issue follows in Chapter 8.

13 *Added value of the Minitools?* We observed that reasoning of high quality only occurred during the lessons without computers, across all teaching experiments. A question that rose was whether using the Minitools had any added value compared to using no computer software. We assumed that experimenting with the Minitools formed a good experience to reflect upon and functioned as a basis for making graphs. The clearest examples of this are the episodes in which students reasoned about Mike and Emily's graphs, which were influenced by working

21. This issue has not been discussed earlier in this chapter.

with the Minitools. Moreover, visually estimating means was afforded by the value tool. Additionally, students were motivated to use the Minitools. We cite from the field notes: "The students liked working with the applets via the Internet." To gain more insight into the advantages of working with such computer tools it would be interesting to compare a teaching experiment in which the Minitools (or similar tools) are used with a teaching experiment in which they are not used.

If we compare these thirteen items with the HLT outline presented at the end of Chapter 5, we can conclude that it has not changed drastically, but has been refined and extended. In short, the estimation and battery activities led to reasoning about many statistical issues including the reinvention of measures of center. To avoid a focus on individual data points, we asked students to work without data and to invent their own data and graphs according to aggregate features. Students in class 1E, having more lessons than the other classes, were able to reason about distribution in an informal way. Apparently, predictions about hypothetical situations helped to foster an aggregate view on data. The episodes in class 1E about distributions, however, were exceptions when compared to the whole set of episodes throughout the year. Concerning the Minitools we had left the initial linear path and were looking for more flexible ways of using them. For instance, we let students compare one data set in the two tools, inspired by theories on multiple representations, and we gave students the freedom to use either tool for solving a statistical problem.
If we compare the statistics unit we developed with the traditional Dutch statistics curriculum, one of the striking things is that we did not teach techniques were only meant for application later. Instead we let them develop their own techniques and notions which at the same time led to conventional graphs and notions as part of performing data analysis. Furthermore, Dutch teachers probably consider the set of traditionally taught techniques and graphs relatively easy. In contrast, the statistical problems we presented to students were complicated from the outset, which urged students to reason about the major key concepts of statistics from the very start. Our approach is therefore more authentic in keeping with the statistical practice of problem solving but also more demanding for teachers. Besides this, most of these problems can be solved at very different levels, which makes them useful for students of different learning abilities and for multiple purposes. The battery problem, for example, was used as one of the first problems, but in a few classes also in the final test with more difficult questions.
Table 6.4 shows the chronological order in which the activities were designed and which activities presented in Chapter 6 led to the HLT tested in Chapter 7.

Table 6.4: Overview of main results that fed the HLT for class 1B

Inter-views	cs, ba, sp												
lesson # class	1	2	3	4	5	6	7	8	9	10	11	12	+
1A	el	ba	bl		-	-	-	-	-	-	-	-	-
1F	el/ bl	ba	cs			wi		sp					-
1E	el	ba	cs	wi	bl			sp			bu	bu	bu
1C	el	ba	cs bl	di	cr	wi		sp		bu		cr	-
1B	el	ba	cs	di	cr	wi	bl	sp	ne	je	bu	te	-

Table 6.5: Abbreviations of Table 6.4

ba	batteries (data analyst role, data creation)
bl	balloon (estimation, sampling, mean)
bu	reasoning with bumps (shape in student graphs)
cr	comparing representations (e.g. Minitools 1 and 2)
cs	compensation strategy for finding mean (with bar graph)
di	data invention (battery context)
el	estimation of the number of elephants in a picture
je	jeans activity (leading to histogram)
ne	new activity (growing samples)
sp	speed trap (shift of hill, quantitative difference)
te	final test
wi	wing span (median, outliers)

7 Testing the hypothetical learning trajectory in grade 7

In the previous chapter we described how we designed a hypothetical learning trajectory (HLT) for grade 7. A notion of distribution as an object had turned out to be difficult to develop, though students in one class were able to reason with shape. The median appeared difficult to develop as a measure of center, and the route planned towards using box plots for describing distributions proved to be time-consuming. We therefore decided to focus on students' notions of spread and sampling in relation to simpler plots such as value-bar graphs and dot plots. Another reason for this focus was that little is known about students' notions of spread, and that our attempts to address sampling had not proven very viable yet.
In the present chapter we test the HLT that was developed for class 1B, a *havo*-class with 23 students (*havo* is higher general secondary education). We first present the results of the pretest to give an impression of the starting point of students' learning process. Next, we compare the HLT specified for particular activities with students' actual learning. In doing so, we analyze how the activities support the development of students' notions of average and spread, and then of sampling and shape (for the method of analysis see Chapter 3). Afterwards, we present the results of the final assessment. In the concluding section, a summary of the results of testing the HLT gives an answer to the first research question of how students with little statistical background can develop a notion of distribution.

Table 7.1: Questions of the pretest

1.	In Figure 7.1a you see two graphs of speeds of cars in km/h. In the Dorpstraat, 60 cars are measured (top graph); in the Stationsweg, 60 cars are measured as well (bottom graph). Each dot represents the speed of a car. a. Do people in one street drive faster than in the other? Explain. b. If there is a difference, could you tell by how much people drive faster in that street? Explain.
2.	After a year, the police of that village have made the graph shown in Figure 7.1b. At the top you see the speeds of cars in the Dorpstraat and at the bottom the speeds in the Stationsweg. a. What can you infer from these graphs? b. What does the height of the graph mean?
3.	In Figure 7.2 you see the average monthly temperatures in de Bilt (in degrees Celsius). a. Can you estimate the average annual temperature without calculation? You are allowed to draw in the graph. b. What is your estimate?
4.	Calculate the mean of the following rows of numbers. a. 6, 7, 7, 7, 9 b. 3, 0, 1, 12

7.1 Pretest

Before the first lesson, students took a 25-minute test (Table 7.1). This test was used to get an impression of their prior knowledge of statistics, which was important information for the HLT. In particular we wanted to find out how students would solve the speed trap problem with dot plots and continuous graphs without preparatory instruction, and how well they could estimate and calculate means. The speed trap activity was taken as exemplary for activities with Minitool 2. We added the continuous sketches to find out if we could use such graphs earlier than in previous versions of the HLT. The rationale for this was that, if students could use such continuous sketches in an earlier stage, we might stimulate an aggregate view on data. Class 1B made the same pretest as class 1C, but performed better.

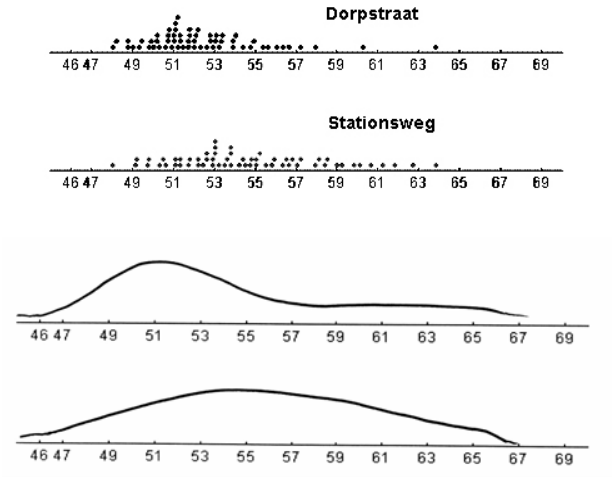


Figure 7.1: Speed plots of questions 1 and 2 of the pretest (unit is km/h)

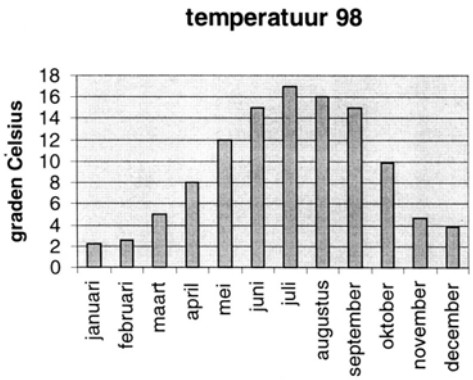


Figure 7.2: Temperature graph of question 3

We discuss the results of the pretest per question.

1a. Of the 23 students, all except 4 wrote that people drove faster in the Stationsweg than in the Dorpstraat, but their explanations differed considerably. One typical correct answer was that there were more dots, or cars, at higher values for the Stations-weg. There were six students who referred to the spread of the data, for example:

> The dots [in the upper graph] are closer to each other.
> The lower dots are more straight.
> In the Stationsweg there is more spread [*verspreiding*[22]] at higher speed.
> The dots are more spread out.
> More dots apart.

Only two students referred to the average ("no, it is roughly equal, on average" and "the dots are further on average"). As in previous classes, we encountered case-oriented views on data, such as "no, the fastest is the same in both."

1b. The question on the difference in km/h was harder to answer. Only five students gave an answer that we consider reasonable (2 and 3 km/h). Many students did not answer the question, which probably indicates they did not understand it. The relatively large variation makes it extra difficult (and perhaps not very sensible) to conceive the mean as a group descriptor.

From the results we concluded that we could use this speed trap activity to let students reason about spread using informal terms, but that it would be demanding for students to quantify the differences and regard means as group descriptors.

2a. Twelve students answered that people in the Stationsweg drove faster and one that people in the Dorpstraat drove faster. Despite the similarity with the dot plots of the previous question (in the eyes of an expert), this continuous sketch was hard to interpret for some students. This is not surprising because students were not acquainted with such graphs without a vertical axis. An example of a wrong interpretation:

> In the first table they only drove fast at the beginning of the street and in the second it is spread out over the road.

This student probably interpreted the axis as representing the street.

2b. Eighteen students wrote that the height of the graph represented "how many there were," although there was no frequency or relative density axis in the sketch. But there were also incorrect interpretations such as "the speed of the cars."

Though the large majority understood that the height of the graph has something to do with "how many there were," we concluded that the interpretation of con-

22. If Dutch terms are unconventional or difficult to translate we also provide the original Dutch ones. *Verspreiding*, e.g., is not a conventional word; it is similar to 'spread-outness'.

tinuous density sketches is not straightforward and should be prepared carefully, for example by paying more attention to density (as opposed to frequency). We dropped the idea of using continuous sketches at an early stage of the HLT. In retrospect, we would not use this pretest question again: without a sense of probability density functions such continuous sketches are likely to be interpreted as frequency distributions.

3a. Most of the answers on how students estimated means were rather cryptic, which shows the need for developing a more precise language in which students can explain how they think about the mean.

> Take the middle.
> Look where the temperature is high, but also low.
> Having an overall look *[Algemeen kijken]*.
> By looking.

And some were incorrect or incomplete:

> Half of the highest temperature.
> Half of the longest bar.
> The lowest is 2 and the highest is about 16 and then half it. [The answer was 7, which means this students probably divided the difference between 16 and 2 by 2.]

When estimating the average annual temperature, three students used a compensation strategy with a horizontal line. During a mini-interview, it turned out one of them had used the midrange.

> I looked at the highest temperature and the lowest temperature, added and divided by two.

When the interviewer asked whether this method would work if only one month had a high temperature, she realized that her method would then yield too high a number.

3b. The estimations ranged from 6.5 to "10 to 12," but most answers were between 8 and 10 (the correct answer is close to 9).
We conjectured that it would be possible to let students reinvent a compensation strategy with bar representations and that many would initially use a midrange strategy as in previous classes. We further assumed that it would be possible to let students realize that taking the midrange is an unreliable method to find an average if the distribution of values is skewed.

4a. Sixteen out of twenty-three students gave the correct answer of 7.2. Of the seven incorrect answers most were 7.1. We assume that the students with the answer 7.1 did not really calculate the mean, but guessed that the answer would be just a little more than 7.

4b. Twenty-one students correctly answered 4 and the two remaining students answered 3.1. These two might have guessed that the answer would be just a little more than 3.

On the basis of the research literature on the mean we had expected that students would have more problems with weighted means and zeros in rows of numbers (e.g. Mokros & Russell, 1995; Strauss & Bichler, 1988), but they seemed reasonably fluent in calculating means, better than the students in Nashville.

From the pretest we concluded that these students, in general, were able to work with dot plots and to calculate means. Certainly not all of the students understood the graphs we had presented them, and from the written test and the mini-interviews we had the impression that the students had not yet developed a suitable language to express their ideas. None of the students talked or wrote about global shapes such as hills. Some reasoned about spread aspects, which confirmed the decision to focus on spread instead of distribution. From the results we concluded that the HLT was tuned well enough to the level of the students and they did not already know what we wanted them to learn.

7.2 Average box in elephant estimation

7.2.1 HLT for lesson 1

In all previous seventh-grade teaching experiments, the elephant estimation had served well as a starting point of the HLT (6.2). In short, it was used to let students think about the variation, density, and distribution of the elephants in the picture, and about a strategy to find a reliable total number, preferably with an 'average box.' This average box is typical in the sense that it contains a typical number of elephants and is normally crowded (in student language: "not too full or too empty"). From the earlier interviews and experiments it appeared that students had different views on average (*gemiddelde*), which we expected to encounter again in this *havo*-class in the first lesson. To stimulate discussions on qualitative aspects of the mean, we again discouraged calculation of the mean. In the HLT, we aimed at evoking the compensation strategy for visually estimating means to make a connection between students' knowledge of the mean algorithm and their qualitative insights into the mean. Furthermore, we expected that such a compensation view of the mean could aid the insight that the mean accounts for all data in the data set. As in class 1F, we would ask what would have been an average box, in this case while showing a matrix with a skewed distribution to question the midrange as a strategy to find a total number (6.6).

7.2.2 Retrospective analysis

During the class discussion on students' different strategies of estimating the total number of elephants, the same strategies were applied as in earlier teaching experiments (Section 6.2): estimating groups and adding the numbers, estimating a number of groups of a fixed size, the area method, and the 'average box' method. The last was used most frequently. This implies that there is a pattern in the strategies that

students came up with in the four different teaching experiments.
In the retrospective analysis, we focused on the different views of average that the students had and wondered if students already looked at how the data were distributed. From the earlier analyses we differentiated the following views on average values and coded students' sayings accordingly during the retrospective analysis with the computer software MEPA (Erkens, 2001). These codes were consistent with approaches to the average as described by Mokros and Russell (1995). We started with the following codes (a-g) and needed to add one during the analysis (h).

a *Algorithm*. With respect to the average as an algorithm there was only one episode. Ciska said, "You should in fact add it all," when asked how to find "an average box." The discouragement of calculations had apparently worked well.
b *Normality*. Most fragments belong to this category. Students used the term 'average box' in the sense of a normal or typical box, as we had anticipated. Some students even literally called it a normal box (*normaal hokje* or *vakje*).
c *Intermediacy*. When asked what they meant by an average or a normal box, they typically answered "not too much, not too little" or "somewhere in the middle" (cf. 5.3).
d *Compensation*. As we had anticipated in the HLT, students started to look at how the numbers in the different boxes varied. Ciska and Susan, for example, were aware of the fact that choosing a full box as an average box would give a total number that was too high.
e *Midrange*. Two students used the midrange: they averaged the number of elephants in the fullest and the emptiest box.
f *Median*. We found no clear example of a median in the first lesson.
g *Mode*. We found no indication of an average as a mode in this context.

In this class there was also a new view on average (h). One group of four students explained that they had roughly averaged their results to get a reliable number. In doing that they probably tried to reduce errors. They explained:

> Well, Babette had something like 270 and Tim had 210 and you had [inaudible] and I had 250 or so.
> Everybody got something different, different number of herds and we averaged that.
> We took about the middle and what was logical.
> Well, what was in between.
> And which was drawn the neatest.
> What was the most reliable [240].

In other words, they chose a value in the center that seemed reliable, logical, or neatest. This last view on average is interesting with respect to the history of the mean, because Greek scientists also chose values that looked reliable (4.5.1).
At the end of the class discussion, students agreed that the method of using average boxes was the most reliable one. The analysis shows that students' notions of aver-

age were much richer than just the algorithm aspect, and that the elephant estimation task helped to bring these views to the fore.
After discussing the elephant strategies, the teacher said that something else was estimated; a picture was divided into boxes and the students had to explain what would have been an 'average box' in this case of Table 7.2. Again, students took different approaches.

Table 7.2: What would have been an 'average box'?

35	58	91
93	83	89
98	97	68
76	82	11

1 *Compensation*. Ciska took a compensation approach to the average box:

> I would take 76, because you have a lot above it, like 97, 98, and 89, but sometimes you are below it, and if you take the upper ones off and add it to the lower ones, then I think 70 is the closest to the mean. (The mean is 72.)

2 *Median?* Ellen then explained, "I looked how many were above it... six were above it." We wondered if she used something like a median, but during an interview after the class discussion she said something different. She started with something that looks like an inward counting strategy but she was apparently aware of the differences in how much the numbers differed from a middle value:

> I just look first at the largest and smallest and those are 93 and 11, or 98 and 11; then there are a few numbers, 93 and 91, and so on, but there are also a few under, 35 and 58, so then I thought that 68 would be roughly in between. (...) I have seen that 76 is not far away and neither is 82, so I thought 68 is roughly in between.

3 *Midrange*. John used a midrange strategy: "11 + 98 = 109, so half of it is about 58." Nico opposed that there were more numbers in the nineties and fewer low numbers: "They are not equal numbers, you cannot compare them like that" (cf. 6.6). He correctly remarked that John's estimation would be too low. Both he and Ciska seemed to be aware of the skewed distribution of these numbers, and this probably held true for more students.
With respect to variation we note that students clearly acknowledged the variation in the number of elephants in each box. In the retrospective analysis we found only one episode that could be a counterexample. Rob said, "All boxes may have about the same number," but this does not mean that he thought that all boxes contained *exactly* the same number; he probably expected little variation.

The experiences in 1B confirm the anticipations of the HLT, because roughly the

same strategies were used and similar views on the average were held as in earlier experiments. Thus, despite the differences between classes, a general pattern in students' reasoning became visible in the different teaching experiments. The power of the estimation activities is that students need to choose an average value and look how numbers are distributed in order to give a good estimate. The apparent purpose is to find the total number, but the hidden agenda is to let students develop notions of average and spread as tools in their reasoning. From the very start students indeed dealt with center, density, variation, extremes, where the majority is, and how the estimated total number is influenced by the average chosen. This was all at a referential level but, as the following sections show, this activity formed a basis for more general reasoning in later lessons as well (as in Section 6.6).
We concluded that the elephant estimation activity was indeed a useful starting point for the HLT. It built upon what students already knew and functioned as a means of supporting students' learning of average values in a more qualitative and coherent way than just applying the algorithm of the mean. In terms of the didactical phenomenology, students indeed organized the estimation phenomena with those conceptual tools that we wanted them to reinvent. In particular, students started to see how the average was influenced by the way in which numbers were distributed.

7.3 Reliability of battery brands

7.3.1 HLT for lesson 2

The general rationale of the HLT was that students would learn to reason about increasingly sophisticated aspects of distribution in relation to shape in increasingly precise ways. From the earlier experiments it was apparent that the battery activity was a powerful one because many aspects of distribution can become topics of discussions at a referential level. The two distributions were deliberately chosen differently, one was symmetrical and the other skewed, so that different distribution aspects could become topics of discussion. As with the elephant activity, students need to take the distribution of the data into account to give an answer to the question of which advantages the brands have. We expected that students would use their notions of average as developed in the first lesson in the battery context as well.
As stated in the didactical phenomenology and Chapter 6, sampling also had to become a topic of discussion. In 1C we had already enlarged the battery data set from 10 per brand to 30 for one brand and 35 for the other one. The first reason for doing so was that it would otherwise not make much sense to compare the data set in both Minitools 1 and 2 (in Minitool 2 it is hard to say anything about shape with small data sets). The second reason was that we considered it inconsistent to teach that samples should be large enough while mentioning at the same time that *Consumer Reports* had only measured 10 batteries per brand.[23] Third, we expected that reason-

ing about different sample sizes could also be a starting point for multiplicative reasoning and talking about majorities, because in the Nashville research the notion of the majority of data did not become a topic of discussion until students analyzed data sets with unequal numbers (Cobb, 2002).
Because the *Consumer Reports* context had turned out to be the most suitable one in previous teaching experiments (6.5), we decided to use that context again in 1B. In contrast to earlier teaching experiments, however, we decided to focus more on spread, because we had noticed that students' attention was often distracted by the mean. We refer to this phenomenon as the 'mean distractor' after Streefland's (1991) 'N-distractor' in the context of learning fractions.[24] Following the Nashville research in which students came to reason about consistency of battery brands as a precursor to spread, we decided to focus the class discussion on consistency to explore qualitative and aggregate characteristics of collections of data points.

7.3.2 Retrospective analysis

To illustrate once more that students generally start with a case-oriented view on data, we mention the observation that the sampling issue of the battery problem was not at all obvious. Nico, for example, explained Figure 7.3 (in which the data values were sorted by size and color) by saying that "you can see the battery fall back." He probably interpreted the data values as the life span values during an experiment on one battery. If so, he took a case-oriented view on the data.

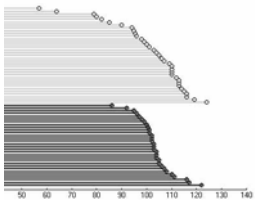


Figure 7.3: "You can see the battery fall back."

In arguing about the brands, the students in previous experiments developed different methods of comparing data sets. The teacher wrote their strategies on the blackboard; they used the mean, the midrange, and thought about the median without using conventional words for midrange or median. This means that they used notions of average they had developed in the first lesson, as expected. Students reasoned with range, outliers, the majority and so on, all situated in a context that was mean-

23. From a brochure of a battery brand we learned that testing batteries is a complex process in which many batteries are tested in realistic intervals and in different devices.
24. When students learn about fractions, many of them think, for instance, that 1/2 is smaller than 1/3 because 2 is smaller than 3.

ingful to them.
When we confronted students with the unequal sample sizes, they found the comparison unfair: the sizes should be equal. However, if we asked during mini-interviews whether one could compare a sample of 400 with one of 405 battery life spans, they thought it was fair enough. This means that they probably did not see the general shape of the samples as saying something about the brands (populations). We return to the battery problem in Section 7.5 on the fourth lesson.

7.4 Compensation strategy for the mean

7.4.1 HLT for lesson 3

According to the HLT, the value of the compensation strategy for finding the mean is mainly that students need to look at how the data values are distributed in order to estimate the mean. As in earlier activities, the hidden agenda of the activities of the present lesson was to let students look at how data values were distributed. Moreover, the mean had to account for all data and become a group feature of a data set. In the third lesson, we combined the context of the elephants and the representation in Minitool 1 to provoke a discussion on finding the mean, as in class 1F (Section 6.6). We expected that the students would further develop their understanding of qualitative aspects of the mean together with the computational aspect. In addition to the HLT of previous teaching experiments, we asked the teacher to do some 'backing' (cf. Cobb, 1999), and ask students why they thought their method worked well. This implies that we felt the need to include advice for the teacher in the HLT (cf. Klaassen's, 1995, notion of scenario).

7.4.2 Retrospective analysis

In the third lesson, the teacher first recalled the elephant task. Showing a slide of the elephant numbers in the eight boxes, represented in Minitool 1, she asked what would have been an average box (Figure 7.4). The students had no trouble in understanding the question; they immediately said 40, 42.5, "41 point something," and 45. When explaining his answer, David drew a vertical line on the slide at 40 and cut off what was too much on the right and added it on the left (Figure 7.4). Other students commented that he "was too low." Apparently, David used something similar to the value tool of Minitool 1 which had been used in the second lesson, and other students understood what he was doing; the vertical value tool had become a tool in their reasoning about the mean (this had not happened in the Nashville research). Another promising issue is that the students did not stick to the actual numbers of the bars (40 and 45), but ended up with the conclusion that the mean would be somewhere in between (42, which is indeed the correct mean as 336 : 8 = 42). This means that students did not think that the mean needed to be one of the actual numbers, in contrast

to what middle school students sometimes think (cf. Mokros & Russell, 1995).

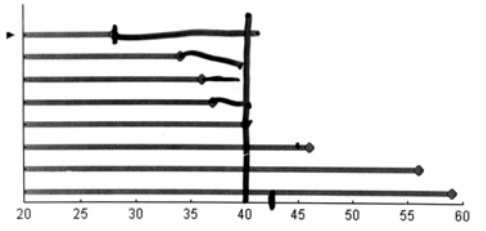


Figure 7.4: David's drawings on a slide

In line with the HLT, the teacher asked why this method of finding the mean was allowed. Ellen answered:

> It stays the same. Together all the elephants remain the same number.

Ciska took a calculation approach to prove that "you get the same mean: you add them all and divide by the number, so the mean is the same." Nico then remarked that calculations were not allowed. In his eyes, she apparently violated the norm that calculations were not allowed.
This whole discussion took place in the first ten minutes of the lesson, which indicates that the compensation strategy was more easily developed than in previous teaching experiments. This could be due to the way we designed the lesson, but also to the similar item on temperature in the pretest or to the fact that these students had more experience with the mean than students in earlier experiments.
In the past we felt the students had not practised new methods enough. To make sure that all of the students would understand and practise the compensation strategy, it was applied in another situation, the battery problem. From the analysis we concluded that most students were able to use this strategy sensibly themselves. If we compare this result with the pretest (question 3) we can conclude that they had learned something new about the mean: in the pretest only a few students estimated the mean visually and no student could explain clearly what he or she did.

7.5 Students' notions of spread in the battery context

7.5.1 HLT for lesson 4

The fourth lesson built on the second lesson, in which the battery problem was addressed. In the present lesson, we wanted to address center and spread issues at a referential level. The quality of a battery brand can be seen as a combination of the average life span and the consistency of the different batteries of the brand. As in earlier teaching experiments, we expected students to come up with different aspects of

the brands that would be related to center, spread, and distribution. In the fourth lesson, we also decided to let students compare one data set in a value-bar graph of Minitool 1 and a dot plot of Minitool 2, because in earlier experiments comparing different representations of the same data set had turned out useful (Chapter 6). In previous experiments we had gained more insight into students' notions of average than into their notions of spread. In the forthcoming retrospective analysis we focus on students' notions of spread.

7.5.2 Retrospective analysis

As expected, the reasoning about the battery brands was similar to students' reasoning in other classes. For example, working with Minitool 1, Ingrid commented that the values of the green[25] brand (K) that "are spread out, there are bad ones and good ones" and "the pinks (D) all have about the same quality." Ho Shan said that "green has more spread" and Tim that "they are much further apart." Students generally agreed that less spread was better.

A short class discussion was instigated to stimulate students to consider more arguments than the ones they had invented themselves. The issues discussed were the following. Some students preferred the green brand (K) because there were many "high ones" and others preferred the pink brand because they were "all in one area." A girl: "Then you are sure that all have that life span." The discussion focused on reliability and predictability, which is probably similar to the notion of consistency that was used in the discussions in Nashville. What was also similar was that some students explained reliability as "at least 80 hours," whereas others interpreted reliability as being within a certain interval (similar observations are made by Sfard, 2000a). There were also indications that students noticed the skewness of brand K: "the greens because it has more at the end."

One mini-interview question that we, including the two assistants, have often asked is, what do you mean by spread? In the following paragraphs we describe how we analyzed students' notions of spread, not only in the battery context but also in other contexts. Thus we exemplify the method of analysis described in Chapter 3: formulating conjectures and testing them at other episodes. Within the battery context, students seemed to view large spread as having big gaps in between dots. The most common description of spread was "how far apart the values are." This led to the conjecture that students interpreted spread as "how far apart the values are" and this

25. As the observations demonstrate, students used color and name indications interchangeably. In earlier teaching experiments we insisted that students used names and not just colors, but due to Nemirovsky and Monk's (2000) notion of fusion we realized that students mostly did not *con*fuse name and color indications. We therefore did not insist anymore that students referred to the life spans of battery brands as 'K' and 'D' as long as their reasoning about 'pink' and 'green' was really about the life spans of batteries. See Roth (2003) for a discussion of the notions of fusion and transparency.

conjecture was indeed confirmed at the other episodes, also in other contexts such as the speed sign context. This characterization of spread does not specify whether students look at the range or some other aspect of the spread. By analyzing the episodes we could make a distinction between a range and a density view. We have not encountered any episodes in which students considered spread as dispersion from a center. We now discuss the range and density views on spread and give an example of an episode in which both views occurred.

Range view on spread
When clarifying his answer on what spread was, David said, "I had the distance between the longest and the shortest dot." This means that he interpreted spread as range (and there were several more students doing so). To avoid confusion between range and other ways of viewing spread, we decided to let students think about the distinction between range and spread in the next lesson. In other words, we decided to guide their reinvention of spread notions to avoid a situation in which the meaning of the term 'spread' would be identified with range.[26]

Density view on spread
Other students had a local view on spread: "Here it is spread out and there the dots are together" or "the green one as more at the end." We interpret this as a local view on spread: students describe the spread in a particular area. We consider this view a possible precursor to a notion of density, which is the reason we call this a density view on spread. We give an example of an episode in which different views on spread were expressed. It turned out that students also considered spread or variation in different variables.

Example of different views on spread
While students worked in groups on their problem, we interviewed the group of Fenne, Susan, and Ciska to find out what they looked at when thinking of spread. We had made up two data sets with the same range but one set of value bars was less spread out from the center (F) than the other (G), as in Figure 7.5a (the b figure shows what Minitool 1 representation it could be a sketch of). Susan said that the spread was the same, which must be a range view on spread. Fenne thought that G had larger spread than F.

26. In Dutch, the difference between the highest and lowest value is called *spreidingsbreedte*. The English term 'range' is used for both the interval as the difference between the highest and lowest value. The Dutch word *spreiding* can mean spread, variation, or dispersion.

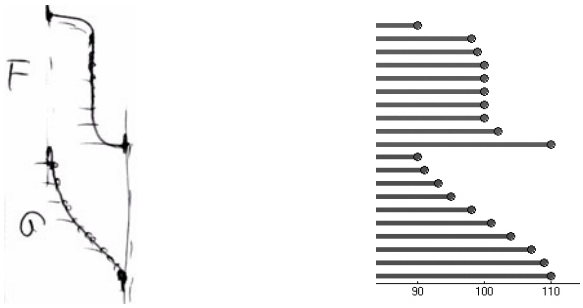


Figure 7.5a and b: Which spread is larger? Sketch (left) of a Minitool 1 representation (right) of imaginary battery brands F and G

Fenne: I think that this (G) has more spread; it stands further apart (*die loopt verder uit elkaar*). And this (F) is here, say from there to here (the middle six values), it is a bit equal and then (the highest values) the spread is large.

Ciska's observations were similar to Fenne's, but she came to a different conclusion. Like Fenne, she seemed to think that there was less spread in the middle part of F and more spread at the extremes, but she concluded that the overall spread of F was greater.

Ciska: Here (in F) it goes back suddenly by a large amount; then there is suddenly a straight part. I think that there is a large spread. (...) Because here (straight part in the center) it is close together and here it shoots out (near the extremes). (...) Yes, then (in G) it goes evenly (*geleidelijk*) and you do not notice that there is a large piece between it (between the extremes) and then I think there is no large spread (in G).

In other episodes we had already seen that students tend to look at differences between data values (cf. Ben-Zvi & Arcavi, 2001). Ciska probably thought that the spread is large if the differences in the distances vary and that the spread is small if the distances are similar ('even'). Thus the similar distances of G showed small spread in her eyes and the variation of differences of F showed large spread.
It is likely that both Fenne and Ciska looked at differences in density in three different areas of the F graph: low, average, and high values. We can reconcile the different conclusions of Fenne and Ciska if we distinguish between the variables they referred to. Fenne probably related spread to the life span variable, whereas Ciska related spread to the frequency or density variable.[27] For instruction, this implies that we have to be explicit to students about which kind of spread we are talking about, and this is easier if students already have a sense of variable, for instance from alge-

27. We have seen these contrasting views in other instances in grade 7 as well. Moreover, similar confusion between spread in the *x*-variable and variation in the frequency variable also occurs when students learn about variation while using histograms, even at college level (Meletiou & Lee, 2002).

bra courses. One intention of contrasting the hypothetical brands F and G was to find out whether students would interpret spread as dispersion from the center, but we have found no indication for such a dispersion view. Mini-interviews such as the one with Fenne and Ciska made us realize how arbitrary it is from a student's point of view to measure spread from a hypothetical measure of center such as a mean. Developing a compensation strategy for estimating the mean that accounts for all data in the data set is clearly not enough to motivate the convention to measure variation from a center value (as with the standard deviation). This could be different in a context of repeated measurements of an assumed true value (Petrosino et al., 2003).

Spread as a conceptual tool

Students' notions of spread were mostly not well articulated, but they were able to use spread as a reasoning tool. We give an example. At the start of the fourth lesson, the teacher introduced the second Minitool by showing the same battery data set in both Minitools and asking how they could get Minitool 2 from Minitool 1. She asked the students which graph belonged to which brand (the order was reversed, as in Figure 7.6, and the slide only offered one gray tone for both brands). It is of course possible to answer this question by looking at the individual data points or range, but students probably used a density view of spread as a conceptual tool; they argued that because D has smaller spread, D must be the dot plot at the top (the dots in the clump of D are closer together). Linking aspects of two representations can be useful to stimulate the use and development of conceptual tools such as the spread notion (see also Chapter 8).

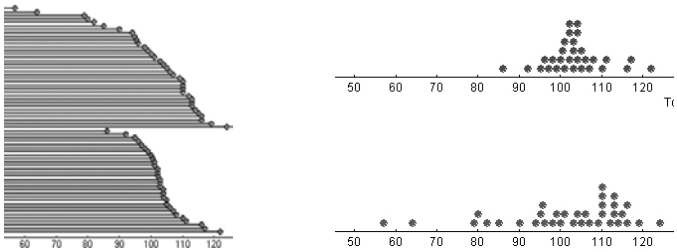


Figure 7.6: Battery problem in Minitool 1 and 2.
Which dot plot corresponds to which brand?

From spread to distribution?

As stated before, we decided to focus on spread (and sampling) in this teaching experiment because distribution would be too demanding. However, we noticed after the fourth lesson that when students reasoned about spread, they often looked at how the data were distributed, for instance, where the dots were spread out and where they were together. In other words, they looked at the density of the data (cf. mini-

interview with Fenne and Ciska). We conjectured that if we let students describe how data are spread out, they would in fact describe how the data are distributed. We also conjectured that it would be helpful to use single distributions instead of comparing distributions, because students tend to compare in vertical slices when comparing two distributions, while they might compare horizontally within one distribution if there is just one distribution to describe. Another reason for avoiding too many comparisons of two distributions was that it seemed dull if every statistical problem involved a comparison (cf. P6 and Cobb & Hodges, 2002).
We concluded that students' notions of spread were strongly linked to the context, and this context presumably made it easier for students to interpret the term 'spread' in a meaningful way as reliability or predictability in relation to relatively simple plots such as a value-bar graph or a dot plot. The next step would be to generalize to other contexts and other representations, and to make distinctions in context-bound notions of spread in range and more sophisticated measures of spread such as the interquartile range (prepared by the four equal groups option in Minitool 2).

7.6 Data invention

7.6.1 HLT for lesson 4 continued

To support students' further development of spread, we decided to use the battery context again and stimulate their interpreting spread as an indication for the reliability of a battery brand. As mentioned earlier, we would introduce the notion of range to avoid spread from simply coming to mean range. We expected that the notion of range would not be hard to understand.
As in 1C (6.7), we decided to let students invent their own data sets according to certain spread-related notions such as reliability. This activity was inspired by different theories that assume a reflexive relationship between symbol and meaning development (Meira, 1995, p. 270) or plead for a back-and-forth movement between different sign systems (Sfard, 2000b). Instead of just interpreting graphs in a context, students should, in our view, also think about how particular features in a context translate into a new graph. If the context were to change, how would the graph change? As with the design heuristic on asking about aggregate properties (6.7), we assumed that students need and develop conceptual tools to link distribution aspects in different representations (see also Chapter 8). The last part of the task was to invent their own data sets that showed certain characteristics:

- Brand A is good and reliable;
- Brand B is good but unreliable;
- Brand C has about the same spread as brand A, but it is the worst of all brands.

In 1C, several students had invented sensible data sets, so we expected similar results in this class.

7.6.2 Retrospective analysis

As expected, the introduction of the range definition was not problematic. We concluded that students interpreted range as intended and spread as how far apart data were in different spots of the graphs. This implies that their notion of spread was local and could be seen as a precursor to density.

The students seemed to enjoy inventing these data sets, but their inventions did not always correspond with what we intended. Two boys interpreted 'worst' as the largest spread and 'reliable' as having high values, the opposite of what we had in mind. This made us realize how arbitrary, in some students' eyes, the links are between good and average on the one hand and reliable and spread on the other hand. One way to improve this task is to replace 'good' with 'having a long life span'. The closest to what we aimed at was the following explanation.

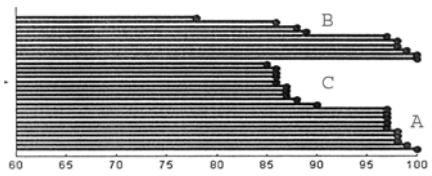


Figure 7.7: Invented data with the explanation:

> Why is brand A better. Because it lives long. And it has little spread.
> Brand B is good but unreliable. Because it has much spread. But it lives long.
> Brand C has little spread but the life span is not very long.

From the retrospective analysis, we concluded that the design heuristic of letting students invent their own data sets according to particular conceptual characteristics can indeed help support an aggregate and conceptual view on data values. As anticipated, most of the students made the distribution of brand C similar to that of A, and not just the spread. This supports the conjecture that students' notions of spread (in a density view) could be seen as a precursor of a notion of distribution ('how data are distributed').

7.7 Estimating the mean with the median

7.7.1 HLT for lesson 5

In the first HLTs, the four-equal-groups option in Minitool 2 was meant to give students a way to measure the center with the median and quantify spread with quartiles. This would further be a prerequisite for characterizing distributions with five-number summaries or even with box plots. In classes 1F and 1E, however, we had noticed that the median was not easy to develop as a measure of center (cf. R16).

Before motivating the HLT for lesson 5, we summarize the problems with the median we encountered in previous classes.
First, many students mistook the median for the midrange, even though they had learned the median as the value that makes two equal groups in Minitool 2.
Second, even when students knew that the median was the value that split the data set into two equal groups, they could rarely find the median in a row of numbers. In hindsight, this is not surprising, because they had learned about the median in a different graphical system, Minitool 2, with the two-equal-groups option and not as the middle-most value (or mean of two middle-most values) in a row of numbers. Because of this one-sidedness in the original HLT and because students probably find tables with data values easier to read than plots (H23), we decided to introduce the median in 1B in the context of values and then link this to the median in plots.
Third, students in the teaching experiments tended to see the mean as more precise than the median, since the latter does not account for how far apart the data are. This last point was already clear in the exploratory interviews (5.1), and from earlier experiences with the average box grids. This view is understandable: to estimate a total, one needs to take deviations from an average value into account. To encourage the use of the median, we assumed we needed to stress its ease of calculation or its robustness, or delay the median's introduction until students have a notion of skewness (Chapter 4: H19 and H20).
Fourth, when students knew the difference between mean and median, they interpreted them as very different things, not as two possible measures of a center. During the final interviews with a couple of students of 1F, we noticed that even in symmetrical distributions they were not willing to estimate the mean with the two equal groups option, because "the mean and median are different things." This implies that they did not take the distribution into account when arguing about mean and median, and it also exemplifies that students can only sensibly use the mean and median if they take the distribution into account (cf. Zawojewski & Shaughnessy, 2000). As in prior activities, we therefore searched for further situations in which students needed to look at how values were distributed to solve particular problems.
For the HLT of 1B, we decided to take a different approach to the median than before: we started with rows of numbers before turning to Minitool 1 or 2 with the wing span data. We focused on three aspects of the median: its ease of calculation, its robustness, and its use in irregular distributions (4.5).[28] For homework, students had to estimate means of rows of numbers within a few seconds. They also had to find out when the middle-most value could be taken as a quick estimate. As in earlier activities, this activity was meant to let students take the whole distribution of numbers into account, but now on a more formal level (with only numbers as the context). We

28. Note, however, that we had not yet completed the historical phenomenology of the median at this point.

assumed that the hint to give a quick estimate would demand a global look at the distribution of numbers and would give the task a playful twist.

7.7.2 Retrospective analysis

In the fifth lesson, the teacher first discussed the homework students had done. Most of the students were able to reason why the median was higher or lower than the mean. However, we also commented in our field notes that part of the discussion was too academic for many of the students. Although students might recall the compensation tasks, the step to a context of just numbers was presumably too big. For the students, there was probably no apparent reason to be interested in the mean or median. Yet there were indications that students were increasingly aware of how numbers were distributed, possibly by the discussion on mean and median. Consider this an example.

The teacher linked the representation of rows of numbers to the representations of the Minitools. This was her own idea; the HLT did not explicitly require this. The question she referred to was:

> Is the median a good estimate of the mean in the following row?
> 33, 56, 89, 118, 120, 151, 163, 165, 165, 177, 178, 181, 182.
>
> Teacher: Assume you had these last numbers in a Minitool, with the dots. Can you imagine where they would be? What it would look like, the dots? Could you tell me?
>
> John: Well, very far apart in any case. It looks a bit as the uh, the green ones with the uh. (From his further explanation it was clear that he meant the 'green' battery brand K in Minitool 1, which was indeed skewed to the left.)

This is an example of successfully linking representations (for other examples see 6.11 and 7.5). By using an aspect of the distribution, in this case the density at different spots of the distribution, John was able to link the big steps in the beginning of the row of numbers to the big steps in the beginning of the value-bar graph, and the high density towards the higher numbers to the corresponding part in the value-bar graph.

In other words, John looked at how the values were distributed. In student language, there are "large steps in the beginning and small steps at the end," and in statistical language this is "skewed to the left." In this fifth lesson, however, there seemed to be few students who understood this link between the two representations.

From such examples we formulated a design heuristic on linking two representations.

> Create opportunities for students to link multiple representations and create a need for using or developing conceptual tools to make this link. Ideally, making this link is a way to solve another question, like which data set is which. Developing such conceptual tools should of course be in the proximal zone of development of the majority of the students.

In retrospect, we concluded that seventh-grade students need much more time to develop a notion of center before they can use formal measures such as mean or median. We consider the informal notions of clump, cluster, and majority as precursors to such a notion of center (cf. Cobb, McClain, & Gravemeijer, 2003; Konold et al., 2002). The historical phenomenology of the median shows what the main reasons were for preferring a median to a mean, but these reasons are beyond the scope of seventh graders with hardly any statistical background. For instance, the median's ease of calculation can only be appreciated if it is seen as one of the ways of indicating a center. To understand the median's use for ordinal data students need to know what ordinal data are. To appreciate the median's robustness students need to understand the influence of outliers on measures of center. To understand the median's use in skewed distributions or for irregular data, students need to have a notion of such distributions (this is the route Cobb, McClain, & Gravemeijer, 2003, propose).
It is of course not very difficult to learn how to find the median of a series of numbers. However, our research shows that we cannot expect that students will then be able to use the median in statistical reasoning, for instance for comparing two distributions. One of the crucial problems students appear to have with the median is that it seems counterintuitive with respect to rational data represented along a ratio scale. The challenge here is to forget about the values of the deviations and only consider the order of the data points. It is not surprising that students say that "the mean is more precise" if we take into account their experience with estimating total numbers. Retrospectively, we hypothesize that students would not find the median counterintuitive with respect to representations in which data points are represented in an ordinal way, such as a series of numbers, dots, or vertical value bars (cf. Bakker, 2002). More generally, there are indications that the median's difficulty heavily depends on the graphical representations used. The position of the median in a new HLT requires further investigation. One of the ideas that seem viable is to address skewed shapes and then describe how skewed distribution shapes are using a median. This route, however, requires considerable investment in first developing a notion of shape.

7.8 Average and sampling in balloon context

7.8.1 HLT for lesson 6

The first five lessons were mainly devoted to center and spread, but from the sixth lesson onwards we focused on sampling and shape. As before, we asked: how many seventh graders would be allowed in a balloon if normally eight adults are allowed? This balloon question builds on the elephant estimation task, but the sampling issue is more difficult. In the elephant task the population is the whole herd visible in the picture, but in the weight context students need to use their context knowledge or

simple sampling techniques to find out about student and adult weights. For the balloon problem we assumed that students would use the same two strategies as in earlier classes (Chapter 6). This time we would use the balloon question to organize a discussion on sampling and students' own graphs (as in class 1E).

7.8.2 Retrospective analysis

In the discussion on the reliability of students' methods of estimating averages, students first focused on the average itself. Tim thought that "the average student is about 45 kilo." A girl said that she included her own weight and that of others in her estimation; exactly how was not clear. It took some effort on the part of the teacher to get the students to think about the *reliability* of their answers. They just seemed to assume that their own estimations were good, and were not inclined to make this abstraction step towards thinking about methods of sampling. Students either relied on small samples or thought that none of the estimates were reliable. For instance:

> You never know for sure. Yes, if you could look at all students from the Netherlands, then it could be a completely different number.

After some discussion on the reliability of the method, Ciska proposed to weigh two boys and two girls, and later another student proposed to measure all students of the class. Ciska argued against taking the whole class: "Two students are missing." She probably preferred a stratified sample with as many boys as girls to a larger sample in which two students could be missing (H11, H12). After some discussion, all of the students then weighed themselves, while the rest worked on a revised version of the wing span activity (6.9). The discussion illustrated that most of the students were small samplers in this problem situation (Watson & Moritz, 2000), and that they still had much to learn about sampling and representativeness.

7.9 towards shape by growing a sample

7.9.1 HLT for lesson 7

In previous teaching experiments the balloon activity was used as a follow-up of the elephant estimation task. However, in class 1B, we wanted to use it as a starting point for reasoning about sampling and shape. For homework students had to make a graph for the balloon driver with which she could decide how many seventh graders she could take in the basket. We decided to start the seventh lesson with students' own ideas, a sample of two boys and two girls: we would then show the data of the whole class and stimulate a discussion on different distribution aspects, focussing on shape in particular. In general, students found it hard to say something about the shape of graphs. Their language for doing so was not well developed; for example, they could not actively use the term 'symmetrical.' During one of the battery activ-

ities, Nico described the shape as "There is a kind of rhythm in it." From the rest of the mini-interview we inferred that he meant that the shape was symmetrical. Other students called such shapes 'even' [*gelijk*].
From the experiment in 1E we inferred that conceptual reasoning about hypothetical situations, "staying away from data," could stimulate an aggregate view on data. We therefore decided to let students predict the shapes of graphs for larger and larger samples, hoping they would feel the need to make continuous sketches instead of drawing many bars or dots. Moreover, we conjectured that reasoning about larger and larger samples would be a good basis for talking about samples versus populations at a later stage.

7.9.2 Retrospective analysis

One class

In the seventh lesson, the teacher started the discussion with the question of why the weight data were created. The first reactions were "to calculate the mean" and "to know precisely." These reactions exemplify the mean distractor effect that we tried to avoid. Perhaps the context did not have any intrinsic motivation for looking at the spread of the weight data. In the balloon activity students had to reduce variation to answer the question of how many students could go into the balloon. The discussion on spread might have been unexpected from a student perspective. If we were to do this again, we would probably design a different activity in which spread would be more relevant to avoid this problem. The teacher then asked why they started with four students and then decided to measure the whole class.

Aster: Yes, because those four could be lying out [*die vier kunnen misschien wel heel erg uitwijken*].

Fenne: There are always kids in the class who are heavier or much lighter. (...) with the whole class you have a clearer average.

The students had made different types of graphs (e.g. Figure 7.8). Bar graphs were favored and, remarkably enough, many students used vertical bars, although Minitool 1 only offers horizontal bars (the same happened in grade 8). Despite their experience with Minitool 1, these students were probably more acquainted with vertical bar graphs than with either horizontal bar graphs or dot plots.[29] In our view, it is therefore an unnecessary restriction that Minitool 1 only provides horizontal value bars. The discussions revealed a few things about the students' notions of sampling. In the previous lesson, Ciska had opposed measuring the whole class, because two students were missing.

29. This could well be a more general phenomenon: in a study by Baker, Corbett, and Koedinger (2002), American eighth and ninth graders mostly drew vertical value-bar graphs when asked to make a histogram of a data set.

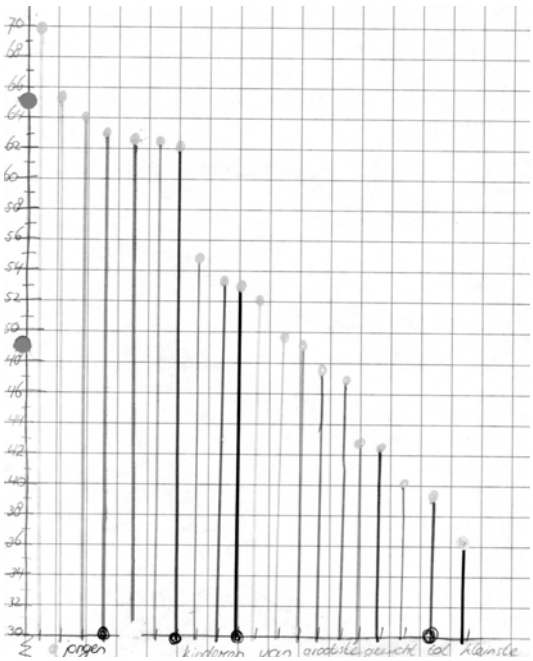


Figure 7.8: Susan's graph for the balloon driver: the sample of four is indicated with black dots; boys and girls' weights with different colors

In the present lesson, she noticed that the data set showed only seven girls, whereas there were nine girls in the class. Aster responded to this:

> I think that this is not so good, because there are more boys than girls and then you are not certain, the ones that are more, maybe these are the outliers.

In the prediction of the whole class, Corinne expected more heavy students, but others expected more students of average weight. When the teacher asked, "What could you say about the larger sample," Tomer observed that "the largest part of the class is around the 50." John then said that "the spread becomes larger." From what he said about it, the teacher inferred that he in fact referred to range and she asked the class:

> Teacher: Do you always look at the lowest and highest?
>
> Froukje: You can also look where the most are. (...) Between 40 and 50.

This is one of the many examples in which a student talks about a group in the middle or the lowest and highest values.

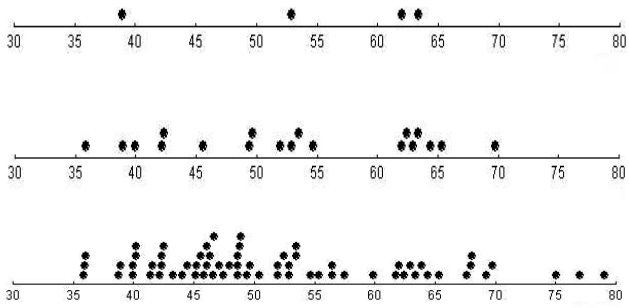


Figure 7.9: Growing weight samples in grade 7 (4, 19, and 67 data points)

Three classes

When the teacher showed the data from other classes, several students reacted with saying, "Wow." They were stunned with the data around 80 kg of one other class. "Who is this?" There were several indications that students were involved in reasoning about the growing samples, and the discussion revealed much about their statistical notions. As in the aforementioned quote, Froukje seemed to look at a modal clump, while Fenne seemed to mean a kind of density when she used the term 'spread' (cf. Section 7.5.3).

> Fenne: They are all a bit around the same weight. And those above, there the spread is much larger, much further apart.

At some point in the discussion we were faced with the problem of having long discussions; students can tire and lose their concentration. For the next time, we decided to let students do more of the growing samples activity by themselves or in small groups (Chapter 9).

Six classes

Next, the students had to predict what the weight data of six instead of three classes might look like in a graph. In general, the shapes of the graphs students made were smoother than those of the real data. Almost all students used dots or small *x*'s instead of bars (see Figure 7.10b and c). We assume that students opted to use dots because we presented them dot plots as feedback and because drawing many bars is rather clumsy.

As a solution to the problem of what to do with very large samples, a few students let one dot signify more than one student. Jeroen, for example, wrote under his graph of six classes:

Because it looks like this with three classes; one dot is just worth more.

Ho Shan indicated the hypothetical averages of six classes called A to F in his graph (Figure 7.10a). Ingrid doubled all 'frequency' values of the three classes; she just added a 35 as a low 'outlier' (Figure 7.10b). Froukje made a shape that is in fact smoother than a typical real data set (Figure 7.10c).

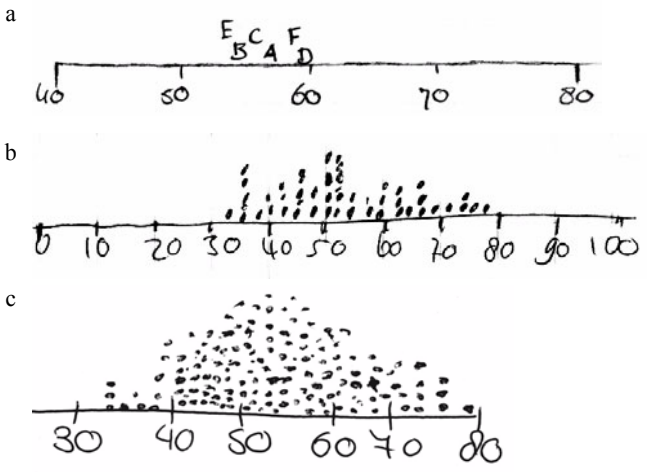


Figure 7.10: Predictions of six classes: a) six averages of Ho Shan, b) Ingrid's dot plot, c) Froukje's dot plot

In contrast to the anticipations of the HLT, students had not made continuous sketches. We therefore used the mini-interviews to ask about a graph of even larger samples. One observation was that if students drew lines, these tended to be not very smooth but rather bumpy. After Tomer had drawn the bumpy line in Figure 7.11, we added three continuous sketches (right-skewed, 'normal', and left-skewed) and asked during a mini-interview:

Int.: Assume now that there are very many relatively light children in the Netherlands and a few that are very heavy, which graph would you (interrupted)?
Tomer: This one. (This is correct: right-skewed)
Int.: And when would you get something like this? (left-skewed)
Tomer: If there are very many heavy students and a few light ones.

Tomer was apparently able to relate the form of the sketch to the meaning (many light students and a few heavy ones). This is an example of a mini-interview that goes beyond what is discussed in the lesson to find out where we could go in the next lesson or a future teaching experiment. In retrospect, we realized that we had implicitly expected continuous smooth shapes. We concluded from the bumpy lines that

students accounted for the variability of the data.[30] Smooth continuous lines are in fact idealizations that do not give a realistic impression of the situation. In other words, where we had tended to think just of the signal, students also took the noise into account and could probably not consider signal and noise as separate constructs (cf. Konold & Pollatsek, 2002).

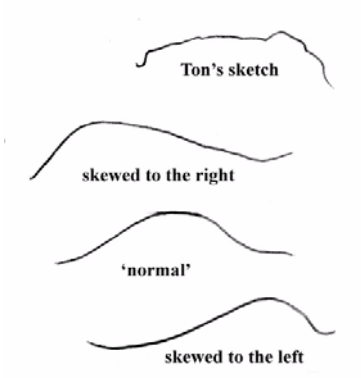


Figure 7.11: Sketches during to mini-interview with Tomer

More than expected, the growing samples activity created opportunities to discuss aspects of distributions. For example, students predicted more values around the average and larger spread if the sample was grown. They had reasoned better about shape than before. We decided to conduct another teaching experiment focusing on distribution in relation to this idea of growing samples, this time in grade 8, when students have a better mathematical background than in grade 7. We envisioned that students, after enough experience with activities such as growing samples, would come to see the stability of distribution aspects such as mean and shape (5.2).

7.10 Average and spread in speed sign activity

7.10.1 HLT for lesson 8

In the Nashville teaching experiment, it was during the speed trap activity around the thirtieth lesson that students started to reason with shape, in this case 'hills.' With the growing samples activity as a background, we expected that the students of 1B would also reason about shapes, although we planned to use the speed activity as early as the eighth lesson. We also decided to use this lesson to find out more about students' preferences for Minitool 1 or 2, and to find out how well they understood the different grouping options in Minitool 2. Again, we slightly changed the phrasing of

30. In American classes of grade 5, 6, and 9 that we have visited, students also made such bumpy or even spiky lines, and never smooth curves.

the problem to avoid vertical slicing and comparing numbers before and after a cutting point, for instance 55 km/h, which is the point where people get fined for speeding. We changed the context to a sign with "you are driving too fast" on it if people drove too fast, and we mentioned nothing about the speed limit in that street. We expected that more students than in the pretest would be able to quantify the difference between the two conditions, because of their developing notions of center and distribution and their experience with the Minitools activities.

7.10.2 Retrospective analysis

In the eighth lesson, the teacher and the students first talked through the data creation for the speed sign. This introduction felt more natural than some others, probably because the teacher did a good job and because students could easily get an image of the situation due to the fact that all knew such signs on roads.

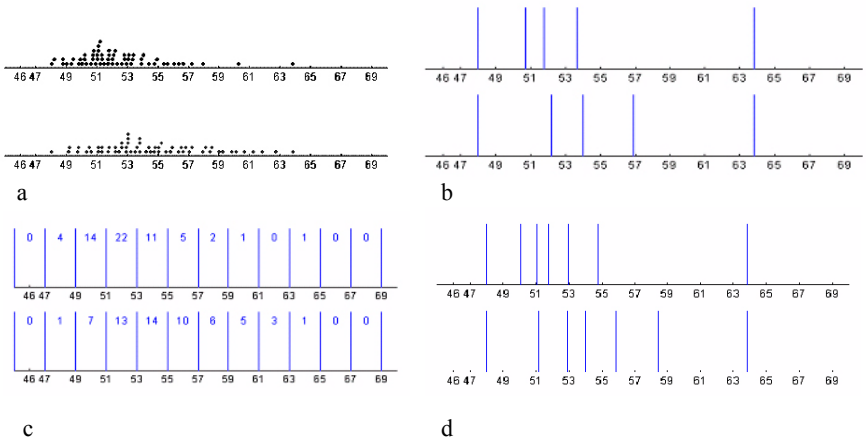


Figure 7.12: a) Speed sign data structured with b) four equal groups, c) equal interval width, and d) fixed group size (10 per group)

Was the speed sign effective? Yes, answered most students, but not by much. One of the questions was how much the effect was in km per hour. Babette and Aster first thought, "how can we know that," but then they immediately started using the value tool for finding means, and found a difference of 1.8 km/h. They concluded that "the largest group is driving less fast." However, certainly not all students had a reasonable answer to this question. Though many students were able to quantify the difference in the two speed situations than in the pretest, we cannot conclude from students' answers whether they interpreted mean and median as group descriptors. Students had to compare the different Minitool options in which the data were hidden. Each representation seems to have its advantages.

From student answers when using the fixed-group-size option, we conjectured that students can see the density in various parts of the graph well when using this fixed-group-size option. For example, Klaas and David had chosen fixed group size 7. Klaas said that the lines were closer together in the after condition. This means "that the spread is less large." David concluded, "so it had effect."
It could well be that the four-equal-groups option helped some students to see where the 'majority' was. For instance, Peter chose four equal groups as the clearest option because "then you see what most do."
We conjecture that reasoning with these multiple representations was useful to grasp the notion of spread. Although students' answers were often vague, it was clear that they started to develop a language to talk about range, spread, and distribution. One of the questions was how the spread had changed. Nico answered that it was "less but with more outliers." Most students by now understood the difference between range and spread, but there were also counterexamples. Jeroen, for example, answered to the question of how the spread had changed: "The spread has not changed but it has shifted." He probably meant that the range was the same, but that the clump (majority, hill, high density part) where "the spread was small" had moved to the left. We interpret this as confirming our conjecture that the differences between spread, density, and distribution are not always clear for students (C7).
The difference between mean and median was not clear to many students. They often thought that the blue line of two equal groups was the average, or they thought that the midrange was the median. Apparently, getting to know the median takes more practice than we had assumed.
The three interviewers asked many students whether they found Minitool 1 or Minitool 2 easier for answering the questions, but the answers were not very useful. One boy had a clear preference for Minitool 2 when looking at spread. Another boy preferred Minitool 1 for small data sets and Minitool 2 for large data sets. Most answers only referred to how easy it was to read off values from the graphs (cf. H23). This means we were not able to confirm or reject our conjecture that students find Minitool 1 easier for estimating means and Minitool 2 for looking at spread. This conjecture probably needs to be tested in a more clinical way.

7.11 Creating plots with small or large spread

7.11.1 HLT for lesson 9

Within the battery context, most students had learned to take the mean as indicating the average life span and the spread as related to the reliability of the brand. This means that they reasoned with spread at a referential level. Later, they learned what range was, and they characterized spread as "how far apart the values are." In this ninth lesson, we wanted to check whether students could invent graphs with large

and small spread at a more general level, without a specific context. We would also ask whether students could make a graph with a large range and a small spread, but expected this to be a tough question.
Additionally, we intended to combine several things that were discussed earlier, such as sample size, shape, and the Minitool representations. For that reason, we asked the students to make an overview of the differences and similarities between the different classes' weights. We had the impression that students liked tasks that ask for differences or similarities. Moreover, the advantages and disadvantages of small samples had to be discussed.

7.11.2 Retrospective analysis

By this ninth lesson, students knew that small samples are easier and cheaper, but less reliable. Almost all of them were able to draw graphs with small or large spread. Some chose value-bar graphs (either horizontal or vertical); others chose dot plots.

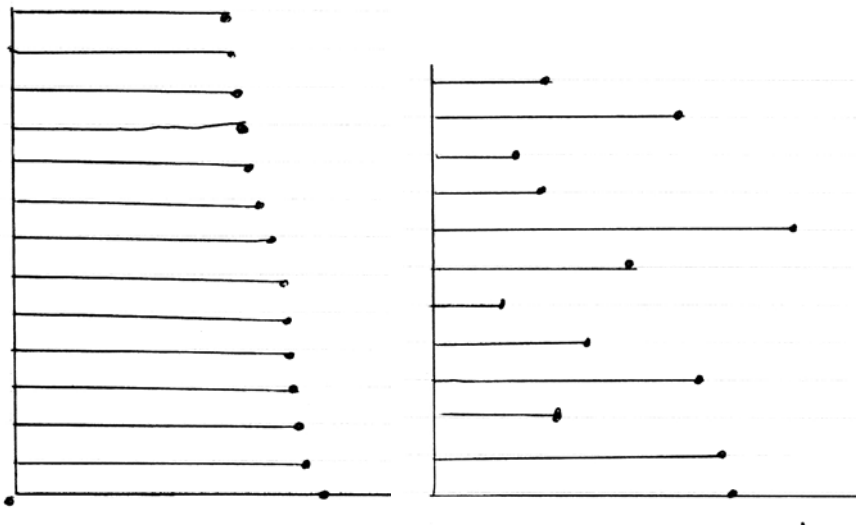


Figure 7.13: Nico's graphs of small (left) and large spread (right).

Figure 7.13 provides one example. As with the data invention activity (7.6), students liked to invent graphs of data sets. During a class discussion we asked what the smallest possible spread would look like. Some students understood that all data points should then be the same ("exactly on top of each other"), but others objected that there would no spread at all. We also asked if students could think of a graph with a large range but small spread. This proved to be a tough question indeed. Only after some help did a few students get the idea and made something like in Figure 7.16.

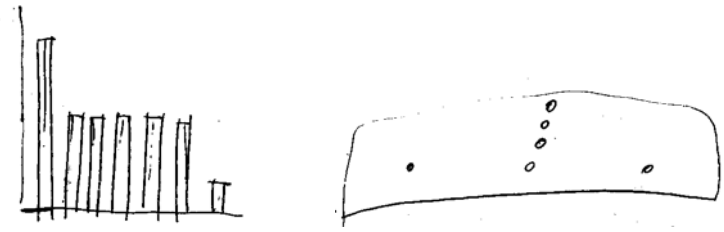


Figure 7.14: Susan's value-bar graph and a dot plot with large range and small spread

We had the impression that students at this stage thought that small spread is best, an idea that probably stemmed from the battery activity. Therefore we asked if students could invent contexts in which large spread is favorable. They mentioned coins (not only 10 cent pieces), soccer matches (different scores are favorable to ties), and fruit (not only apples of the same size). We concluded from this lesson that students had developed a better understanding of samples, and a more general understanding of spread in relation to two different graphical representations, value-bar graphs and dot plots.

7.12 Jeans activity

7.12.1 HLT for lessons 10 and 11

If you know the waist sizes of 200 men, which percentages of each jeans size would you advise the factory to make? We considered this question a way to avoid the mean distractor effect, because the whole distribution is important, not only the mean. Additionally, we expected that this question would offer students a more quantitative and precise way of structuring a distribution than simply working with low, average, and high values.

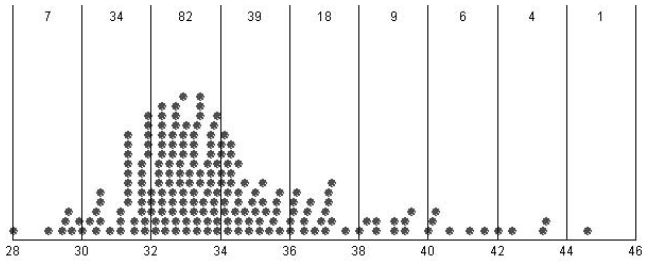


Figure 7.15: Jeans data set with equal interval width

In that sense, this jeans activity built on the speed sign activity and other activities that supported students' notions of spread. Because we had learned from previous teaching experiments that setting the stage for such an activity was very important (talking through the process of data creation), we decided to spend one lesson on the question of how one could investigate the question of the jeans factory, and the questions of who else might be interested in people's measurements, and how one could best take a sample. Moreover, this activity was meant to motivate using the equal-intervals option and support future understanding of the histogram (the revised Minitool 2 with histogram and box plot overlays was ready but, unfortunately, it did not run properly at the school site).

7.12.2 Retrospective analysis

It proved harder than expected to let students invent a design for investigating the question of the jeans factory. Most students were pragmatic: if you have left-overs in one year you know you have to order fewer of that size the next year. As in previous attempts it turned out hard to engage students in discussing sampling issues.
In the next computer hour, many students first chose the four-equal-groups option to structure the given data set. It could well be that they chose four equal groups because the percentages would then be easy to find. Their answers were very general; for example, "make a lot of sizes from 32 to 34 (inches) and a little less of 36 to 46." We then asked if students thought that a factory paying 1000 guilders would be glad with such advice. Many students then looked for more precise ways and eventually used the equal interval width option at 2 inch, which provided frequencies. They had no problems with dividing these numbers by 2 to find percentages. From the mini-interviews and their answers we inferred, however, that hardly any student had understood the details of rounding off or up (a person with waist of 33.2 needs size 34) or of the representativeness of the data set for other groups of men.
The clothing sizes context was too complex and not really appealing: students just try clothes on, they rarely even know their size, and they are usually unfamiliar with what inches are. We concluded that this activity needed revision in several ways. For a next time we would prefer to invest more time in sampling, center, and spread issues than trying to guide the reinvention of a histogram or a box plot, for example. We do not deny that reading off values from a histogram or box plot can be easy (cf. Baker et al., 2002), but in our approach graphs had to stand for distributions and become reasoning tools about aggregate features of the data sets.

7.13 Final test

During the twelfth lesson, students made a final test with four tasks, which are described in this section. Almost all students were also interviewed in pairs on one of the tasks.
This final test was only part of the assessment: their answers to the activities in pre-

vious lessons were also judged on completeness, neatness, and quality of argumentation.

First task: row from small to tall

What would it look like if a random sample of 100 men were to stand in a row from short to tall? Gravemeijer (1998c) asked many people this question and the mistakes and surprises were numerous. More than a century earlier, in 1871, intrigued by Quetelet's work on the normal distribution, Knapp reported on how he drilled recruits. He ordered them to stand in a row by height. Freudenthal (1966a) wrote about this:

> Toen hij tot zijn schrik bemerkte, dat de kruinen niet naar links èn rechts afliepen, dus niet zoiets als een normale kromme aftekenden, liet hij ze heel eigenaardige exercities uitvoeren, waarvan de finesses mij niet duidelijk zijn geworden. (p. 137)
>
> When he noticed to his horror that the line of the tops of their heads did not slope to the left and to the right, i.e. that the line did not look like a normal curve, he let them do very peculiar exercises of which I do not understand the subtleties. (translation from Dutch)

Apparently, it can be hard to imagine the line over the tops of people's heads. These two references inspired us to ask how we could picture the whole class in a row from short to tall. This question was meant to discover whether students could relate their understanding of distribution in dot plots to a value-bar graph or a continuous shape.

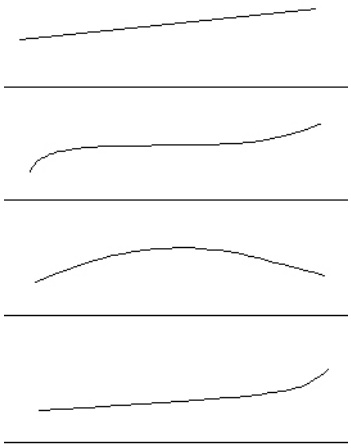


Figure 7.16: Four sketches that were provided to "help or mislead you."

If students could transfer their intuition that there are many people around average into a flat part in the graph, we could then conclude that they had some conceptual

understanding of variation and distribution with which to make this step of transfer. We provided four continuous sketches as a reference context "to help or mislead you," as we wrote in the task.
Three students of the 21 who made the test gave no answer, and three drew straight lines. Six students drew a bumpy continuous line without a horizontal part in the middle, and three used vertical bars or sketches of people of linearly increasing height. We assume that they did not account for differences in density or frequency, but that the six with bumpy lines did account for some variation. Four students drew a line with a horizontal part in the middle and four drew bars or a line of people with a flat part in the middle. From their explanations we inferred that they thought of three groups of low, average, and high values (see C1 in Appendix).

> There are tall children and short children and most of the children are about equally tall [*zijn toch wel even lang*].
> There are a few short ones and a few tall ones and the rest is average.

The fact that students drew sketches without a flat part in them does not mean that they thought that there was no difference in frequencies of certain heights. Some of those who had made straight lines showed insight into this task during the interview. Fenne, for example, explained her sketch as follows:

> But look here, from here to here, look, it is a little higher, this a little lower, but this piece [in the middle], that is roughly about the same. And these are then again the high outliers and these the low outliers.

Because she used 'outliers' in an unconventional way, we asked her peer:

Int.: Do outliers occur often or rarely, Corinne?
C.: Rarely.
Int.: Rarely.
F.: No, I think that [interrupted].
C.: Otherwise they would not be called outliers.
F.: No, I do think that they occur often. I think that if you go to any class here, whether 5 *vwo* or 6 *vwo*, or 2 *havo*, or whatever. You always have.
C.: But not many, because otherwise it wouldn't be outliers.
F.: No, but I think, no, that you'd get that in all classes [*dat je wel bij alle klassen zo uitkomt*]. With animals too, uh and uh with everything, I think.

From Fenne's words, we infer that she had a sense of the stability of distributions. She expected values with a low frequency in all data sets. Fenne interpreted 'outliers' as extreme values of which there are only a few, but Corinne interpreted it as values that are rare or exceptional. Though Fenne was not very good in mathematics or other school subjects, and did not work very well for most of the lessons, she demonstrated a sense of the stability of how data are distributed. In retrospect, we had paid too little attention to culturally accepted ways of using the term 'outliers'.

From the retrospective analysis we concluded that students tended to think of data sets as three groups of low, average, and high values (C1). There was only one exception that we could interpret as a qualification of that observation. Ciska made the sketch in Figure 7.17 and explained it as follows: "There are 3 smaller ones, about 10 average, 3 to 4 taller, and of course in between." It could be that she realized that the numbers she guessed had to add up to 23.

Figure 7.17: "There are 3 smaller ones, about 10 average, 3 to 4 taller, and of course in between."

We concluded that two thirds of the students were able to express their notion of distribution consisting of low, average, and high values in some graphical representation with a continuous line or a series of bars or people. We assume that their experience with both Minitool 1 and 2 was a prerequisite for this ability.
Though this first task was informative about students' notions of distribution, we would not use it again in this form as an assessment task. As the mini-interview with Fenne demonstrates, some students initially chose the wrong shape, but could still explain why there should be an almost horizontal part in the middle of the shape. Without such interviews, we would have had very different findings.

Second task: training program
The second task was meant to see if students could make graphs that were compatible with a context story on running practice with both informal and statistical notions. There were no restrictions on the type of graph they could use. We had deliberately incorporated characteristics in the story that ranged from easy to difficult, so that all of the students could display a number of characteristics in their own graphs on their own level. The easiest feature was case-oriented ("the fastest runner needed 28 minutes") and the more difficult ones were aggregate features such as "the spread of the running times was much smaller than in the beginning but the range was still pretty big."

> A seventh grade class is going to train to run 5 km. To track their progress they want to make three graphs. One before training starts, one halfway through, and one after ten training sessions. Draw graphs that belong to the following story:
>
> 1. Before training started some students were slow and some were already very fast. The fastest ran the 5 km in 28 minutes. The spread between the other students was large. Most of them were on the slow side.

2. Halfway through, the majority of the students ran faster, but the fastest had improved his time only a little bit, as had the slowest.
3. After the training sessions had finished, the spread of the running times was much smaller than in the beginning, but the range was still pretty big. The majority of the students had improved their times by about 5 minutes. There were still a few slow ones, but most of the students had a time that was closer to the fastest runner than in the beginning.

The majority of the students (15) used dot plots, five used bar graphs (four of them vertical and unordered), one made continuous graphs of students' running improvement, and one made a kind of scatterplot of time and distance with the explanation, "some did not reach the 5 km." Figure 7.18 shows one of the better symbolizations of the story.

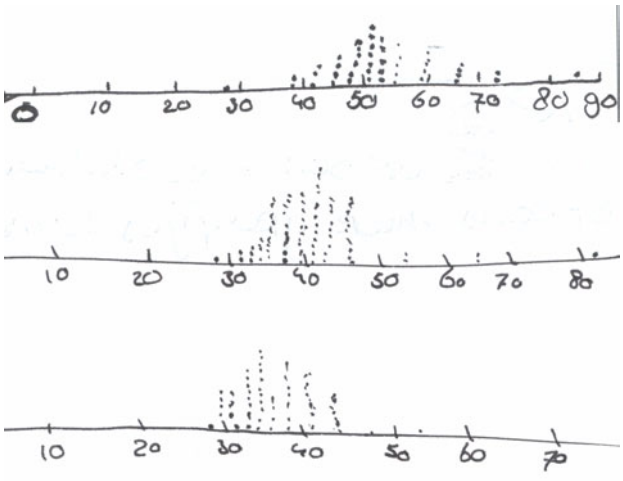


Figure 7.18: John's dot plots for task 2

In general, the students were able to represent certain characteristics such as range and spread, but not all details in the graphs were compatible with the story as can be seen from Table 7.3.

Like the first task, the second revealed much about students' abilities to translate conceptual items into graph characteristics. However, such student graphs are laborious to assess (we gave points for each correctly represented feature). If we take into account that interpreting given graphs is mostly much easier than choosing a suitable graph and representing such aggregate features as shown above (cf. Baker et al., 2002), the results on this task are promising.

Third task: larger samples

The third task was designed to assess students' understanding of samples both on a

basic and an advanced level.
3a. What is an advantage and a disadvantage of a large sample? Advantages that students mentioned were: more precise (5), more certainty (4), good overview, better image, closer to reality, more reliable, then you know it roughly, better average, higher chance that it is correct. Disadvantages they mentioned were: takes much time (11), much work (3), much money, difficult (with many people). Almost all of the students gave sensible answers.

Table 7.3: Answers of 21 students

1. fastest 28 minutes (extreme value)	The large majority (18) indeed represented this correctly. One student only represented this item and nothing else. Two students did not provide an axis with numbers and three (in total) did not correctly represent the situation.
1. most were on slow side	11 students had made a distribution skewed to the left (clump at the higher end), 7 had not done so, and in 3 cases it was unclear to us whether the shape was skewed or not.
2. slowest and fastest slightly improved	14 correct and 7 incorrect
2. majority faster	In 14 cases the majority was indeed faster and in 6 cases it was not. One case was unclear because there were no axes.
3. spread smaller than in first graph, but similar range	Though the large majority used a similar range, only 8 students clearly made the spread smaller by clumping data more together (such as in John's graph).
3. majority 5 minutes faster	11 of the graphs showed a majority that was from 5 to 10 minutes faster.
3. skewed to right	In 7 cases we can see a transition from skewed to the left or pretty symmetrical in the first graph to skewed to the right in the third graph.

3b. What is a representative sample? We had only paid very little attention to the notion of a representative sample, which makes the answers uninformative.
3c. If you grow a sample, how do you know when you need not grow it much larger? We considered this the most difficult question and we had hoped that some students would write that one could stop growing a sample when some pattern occurred or when the shape did not really change anymore. The closest to this was:

> If not much change occurs (*als er niet veel verandering in komt*).

Other answers were:

> There are many of about the same weight.
> Because you double everything at a certain moment (...)
> Take the average of every class.
> If more and more groups come at some spots.

Most of the students, however, did not understand this question. From the mini-interviews we inferred that only a few students understood that they could probably stop once the mean no longer really varied. Nor were there many students able to reason about spread.
From the results we conclude that almost all of the students knew some advantages and disadvantages of large samples, but none of them were really able to reason about shape issues. This is not really surprising after one lesson on growing samples. It shows that developing a notion of shape takes time and careful design.

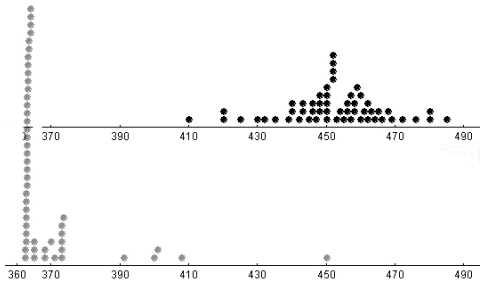


Figure 7.19: Times of two running stages in the Tour de France (not real data)

Fourth task: Tour de France
In a way the fourth task was the reverse of the second task: from two graphs the students had to make up a story that was compatible with the graphs. The data were times in stages of the Tour de France. Two stages were pictured: one with a closed pack of cyclists finishing together and one with spread-out arrival times.

> Make up a small story about the two stages. What kinds of stage were they? Through the mountains or flat? With a leading group or was it one pack? Were the stages tough or light, short or long?

Four students did not answer this last question and two gave an incorrect answer. One student probably did not realize that winners have shorter times ("the first is much faster than the second, but there is one who is much faster than the others"). The other 'incorrect' answer included statistical terms:

> The first is not in the mountains; they are all close to each other and there is a large spread. The second is in the mountains; there was one very good one. There was a large range and small spread.

Fourteen students answered that the first run was in the mountains and the second on flat land, and that the first took a long time and the second a short time. Their reasoning differed. Two explicit examples were:

> Run 1 was probably a mountain stage because the people are not close to each other (long run). Run 2 was probably on flat land because they all stayed close to each other. The first was a mountain stage and the second was on flat land. With the second there was a large pack and they needed less time and the first had one leading group of which one broke away.

From the results we conclude that the majority of the students were able to interpret these graphs within this context. A weak point of this task was that it was not very informative about student notions of spread or distribution, and those students who did not know this context well were probably at a disadvantage.

7.14 Answer to the first research question

In this chapter we set out to answer the first research question for grade 7:

How can students with little statistical background develop a notion of distribution?

The HLT provides an answer to this question. We start with the confirmed conjectures about students' key notions in the beginning of the teaching experiment, the starting points of the HLT (7.14.1). Then we discuss the revisions in the end goal of the HLT (7.14.2), and how the instructional activities of the HLT supported students' understanding of distribution aspects and sampling (7.14.3). Finally, we summarize an answer to this first research question for grade 7.

7.14.1 Starting points of the hypothetical learning trajectory

Variability. In general, students are familiar with variability. They know all too well that not every student in the class has the same height or weight, for instance.

Sampling. The seventh graders had a poor notion of sampling. In line with the findings of Watson and Moritz (2000), many students thought of small samples and did not specify methods of selection. For example, in the weight context they first proposed to measure two boys and two girls (Section 7.8) and some wanted to weigh everyone in the country. In the balloon context, students were inclined to trust their own estimations and not to think about how samples could be taken.

Data. As in previous classes, we encountered many inappropriate case-oriented views on data, especially in the beginning of the teaching experiments. Our findings are consistent with other research that shows that middle school students generally do not see a data set as a group with group features but as a series of individual numbers (e.g. Konold & Higgins, 2003). Furthermore, students still need to learn that claims should be made with the help of data (cf. 5.1).

Center. The results of the pretest confirm the image of the exploratory interviews (5.1): Dutch seventh graders are reasonably fluent in computing means, but their knowledge of the mean as an algorithm appears somewhat separated from their 'average group' intuition or other views on the average (mode, midrange, balance point, median). In the pretest, three students used a compensation strategy for estimating the mean of the annual mean temperature, but this could have been a simple midrange estimation. A visual compensation strategy still had to be learned. In well-known contexts such as height and weight, students seem to have strong intuitions about what is normal and what is not. They use the term 'average' for the group of normal cases. This average group is sometimes called the majority or, with respect to dot plots, a clump (*klont, kluit, hoop*).

Spread. The first way in which students characterize the spread of a data set is by its range (a range view on spread). We assume that many also look at how the other data values lie in between these extreme values, but without statistical notions and graphs it is hard to express a more precise view on spread.

Distribution. In Chapter 6 we mentioned one conjecture on students' views on distribution that was confirmed in the present teaching experiment during the retrospective analysis: students tend to group data, especially imaginary data, into three groups of low, average, and high values (C1). We assume that students already had an intuition about this but could only express this view due to some experience with statistical reasoning with graphs. One simple way in which such a grouping into three or even five is used is in surveys, for example on the service of a hotel as in Figure 7.20.

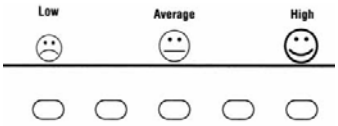


Figure 7.20: From a survey on the service of a hotel

Graphs. Dutch seventh-grade students are acquainted with bar graphs and to a certain extent with line graphs, but not with dot plots, histograms, or box plots. We were struck by the large amount of vertical bar graphs that students made even though Minitool 1 only provides horizontal bar graphs. In our view, this is a reason to include vertical value bars if a new version of Minitool 1 were to be made.

7.14.2 End goal of the hypothetical learning trajectory

In Chapter 5, distribution as an object-like entity is taken as the end goal of the HLT. In line with the Nashville research, the way to achieve this is by focusing on shape.

To a certain extent, this was accomplished in class 1E (a *vwo*-class with fifteen lessons), in which a substantial group of students were able to reason with bumps (shifting or growing bumps, for example). However, in Chapter 6 we concluded that the initial end goal was too ambitious and that spread and sampling should have a more prominent role in the HLT. For the *havo*-class 1B with only twelve lessons we therefore decided to focus on spread instead of distribution and spend more time on sampling. In the teaching experiment described in this chapter, it turned out that by focusing on a density view on spread, we can in fact let students reason about how data values are distributed. When describing how dots were spread out, students mostly described how the data values were distributed.
We assume that it makes sense to let students describe single univariate distributions for the following reason. When describing one distribution, students compare different parts of the distribution in the horizontal direction of the variable ("here it is together and at the end the dots are further apart"). When comparing two distributions, students are inclined to compare vertical slices (e.g., the top graph has 23 in this bin and the bottom one only 17).
During the activity of growing samples, the opportunity to discuss shape arose in a natural way. This implies that even though we focused on spread and sampling, the instructional activities also supported students' development of a notion of shape, albeit not as sophisticated as in class 1E.

7.14.3 Instructional activities of the hypothetical learning trajectory

As in earlier versions of the HLT, the guideline was to support students to reason about distribution aspects with increasingly sophisticated notions and graphs, but we became more and more convinced that sampling needed to be addressed. In the HLT for 1B we initially focused on center and spread, and then tried to draw everything together when discussing growing samples. In this section we discuss patterns in students' learning and how the instructional activities supported, or failed to support, the learning process we aimed at.

Estimation and compensation strategy
The first aim was to develop a notion of center and a sensitivity to how data are distributed. In particular we tried to connect students' knowledge of the mean as an algorithm and informal notions they had of average including a group sense of average. In the elephant estimation task, students used an 'average box' as a multiplicand to find the total number of elephants in the picture. Whilst explaining what they meant by an average box they demonstrated understanding of the several faces of the mean. We used skewed distributions to challenge the midrange strategy and to stimulate students to look at how values were distributed. With respect to the value-bar graph of Minitool 1 students developed a compensation strategy to visually estimate

means. We concluded that students came to see the mean as accounting for all data values and indeed started to look at how those values were distributed. The pattern of students' arguments and strategies in the elephant and compensation tasks were similar to those in previous teaching experiments, and this holds for many of the activities we discussed in this chapter.

Battery life span and data invention

The elephant and compensation tasks mainly aimed at developing a notion of average. With the battery activities we intended to foster a referential notion of spread as reliability or consistency as well as other distribution aspects. In the beginning of the sequence we encountered several examples of case-oriented views. As early as in the second lesson however, students also discussed many distribution aspects using informal terms in relation to the value-bar graph. In the battery activity students indeed came to reason about majority, outliers, how spread out dots (or endpoints of value bars) were in various parts of the graph, and how reliable the brands were. One way in which we guided students' development of spread was by introducing a name for range (*spreidingsbreedte*) to avoid that the term 'spread' would be identified with the term 'range.' In this way, the term 'spread' was reserved for a density view on spread. When using Minitool 2, students had several grouping options at their disposal to support their claims about how spread out data points were, for example with the options to make four equal groups or make groups of a fixed size.

One design heuristic that turned out useful was to let students invent a data set that had certain aggregate characteristics, e.g. an unreliable battery brand with long life spans. In this way, we can stimulate students to think in terms of aggregate features and thus overcome a case-oriented view. At that point many students were able to connect long life spans to average and reliability to spread at a referential level.

Median

The activities around the median were meant to promote students' understanding that not only the mean but also the median can be used to indicate the center of a distribution, for instance for comparing two data sets. In class 1B we tried to take advantage of the ease of calculating medians and let students quickly estimate the average of rows of numbers with medians, but the discussion on this issue turned out to be too academic for them. This implies that our design, like that of the Nashville team, was not very successful on this point.

Can history teach us anything about this? In the historical study of the median, it appeared that the median was born as an alternative to the mean, for instance in cases in which the mean was not robust enough or if the distribution was not symmetrical. One suggestion was to use chance contexts because the median's birth is historically connected to the chance of one half. In hindsight we assume that students need more

time to develop a notion of distribution, outliers, and center as a group, and to get used to the median in different representations such as rows of numbers, value-bar graphs, and dot plots.
One design problem we faced during the retrospective analysis was a conflict between students' notion of distribution as consisting of three groups (low, average, and high values) on the one hand and the median as acquired by two equal groups or the quartiles as provided by the four equal groups option in Minitool 2 on the other. One possible route could be to use students' intuitive grouping into three and design problems in which the median is useful as a one-number summary of the location of the middle group, even in skewed distributions (cf. Konold et al., 2002). At some point there should be the need of a way of quantifying the spread in a conventional way, for instance with four equal groups. The position of the median in a new HLT certainly requires further research. One route that sounds promising is by doing repeated measurements and using the mean or median as an indication of the true value (Konold & Pollatsek, 2002; Petrosino et al., 2003).

Balloon
The balloon activity is similar to the elephant estimation task, but the sampling issue is more complex. In the balloon question, which was inspired by the history of statistics, we tried to link students' reasoning with arithmetic means to their intuitions about sampling. The solution strategies were similar to those in previous classes, and again it turned out that students were not inclined to think about sampling methods. Although we see the use of this balloon activity, we also came to acknowledge a drawback during the 1B teaching experiment: the strategies make use of a mean. This made it more difficult to draw students' attention to other issues such as spread and shape in a later stage. The context of the question might have been partially responsible for the mean distractor effect that we encountered on several occasions, for example during the growing samples activity in the seventh lesson. Many students seemed to think that the goal was to know the mean more precisely, whereas our hidden agenda was to discuss spread and sampling issues. With another problem situation than the balloon question, the mean distractor effect might have been less. However, our impression is that, whatever we do, students tend to use the mean, just because mean and range are the only exact statistical measures they know.

Growing samples
The activity in which we started with small samples that students suggested themselves and in which we let them predict features of larger samples proved more promising than previous attempts to address sampling. The growing samples activity fostered coherent reasoning about many key concepts: when comparing predictions with real data sets of growing size, students reasoned about the center and spread of

the data sets and predicted the shape of even larger samples. The results of this activity were so promising that we decided to test the following conjecture: students can develop a notion of distribution by reasoning about growing samples. The growing samples idea was therefore taken as a recurring theme in the eighth-grade teaching experiment.

Spread

In the pretest, students used predicates such as 'closer together' and 'more spread out' to characterize the difference in the two speed situations. During the teaching experiment, students learned the term 'range' and came to describe spread in a more detailed way (for example, where in the distribution dots were close together or spread out). As the episode in Section 7.5 shows, it is important to specify which type of variation is the topic of discussion. This requires some insight into the role of variables (e.g. life span or frequency?). Moreover, students came to use the Minitool 2 options to make four equal groups or groups of fixed size to compare the spread in subsets of data. In the speed context, they were able to see that the spread was small if lines were close together (Figure 7.12b & d). In the ninth lesson most students were able to make graphs with small or large spread; only a few could make a graph with large range but small spread (Figure 7.16).

Minitools

One of the conjectures that we were able to confirm was that students have no clear preference for one of the Minitools for solving problems (C10). In the previous chapter, we conjectured that students would find Minitool 1 easier for estimating means and Minitool 2 for seeing the spread. Though there were a few mini-interviews that support that conjecture, there were too many other views to confirm it. One such view was Tim's: he preferred Minitool 1 for small data sets and Minitool 2 for large data sets. Most students only referred to how easy it was to read off values from the plots.

We have not found strong evidence that Minitool 1 should be used before Minitool 2 if students can already read Minitool 2. From a design perspective it still makes sense, however, to use Minitool 1 before Minitool 2 (perhaps only for one or two lessons) because the position of the dots in Minitool 2 can be better understood as the endpoints of value bars collapsed down towards the axis. Otherwise it might be unclear why the dots are stacked the way they are (this explanation at least helped a few adults who did not understand why there is no frequency axis in Minitool 2).

Compared to many other statistical applications, the Minitools are very simple. A major advantage of these tools is that students can use them almost without any instruction and use them sensibly in their reasoning about data sets. This implies that there are almost no instrumentation problems, problems with tools becoming instru-

ments in students' mathematical activities (Drijvers, 2003).
It is easy to make a list of features that would make these applications more sophisticated (transitions between bar and dot plots, vertical bars, adjusting the icon size, import of data sets, and so on). The limitation of the present options also limits the variety of what students can do with the data sets. This has two sides. One side is that the teacher can more easily orchestrate class discussions on students' answers than if the software offers many more options (cf. Bakker, 2002). The other side is that the options offered can be too limited for genuine data analysis in the EDA spirit. We conjecture that the seventh graders we worked with could have managed more complex options. One sign of this is that they generally made more vertical value-bar graphs than horizontal ones when asked to draw graphs of situations.
Across all teaching experiments we noticed that students' reasoning was better with no computers present. In a future teaching experiment, we would probably reduce the time for exploration with computer tools to about one third of the time instead of half (Konold, 1995, made a similar observation). This creates more time for reflection.

Final test
From the assessment results we conclude that students had indeed learned important things about sampling, center, spread, and shape in the previous 11 lessons. We summarize the conclusions per task.

1 About two thirds of the students understood that a group of students with average height in row leads to a horizontal part in the image. We consider this a transfer of seeing three groups in dot plots to another representation (similar to 1E, Chapter 6). Apparently, their notion of distribution was not tied to one representation only.

Figure 7.21: Susan's drawing for question one.
Note the flat part in the middle of the row.

2 The most informative assessment question was the one for which students had to make graphs that matched a story on running trainings. There were details in the story that almost every student could represent (the fastest student run in 28 minutes), but also more sophisticated features that only a few students represented correctly (e.g. smaller spread but similar range). If we take into account that other studies show how difficult it can be for students to select and produce appropriate

graphs, the symbolizations of these students were impressive (Baker et al., 2002; Friel et al., 2001).

3 Almost all of the students knew advantages and disadvantages of large and small samples. Not many understood what a representative sample was, but we had not spent much time on that. Furthermore, most students had not come to understand the stability of aggregate features when growing samples. This is not surprising after just one lesson on growing samples.

4 The majority of students were able to interpret the Tour de France graphs. This means they were able to translate the spread of the data into meaning in the context: dots being together signifies a pack of cyclists, which is common in flat stages, whereas dots being spread out signify scattered cyclists, which is common in mountain stages. A disadvantage of this question is that students need context knowledge of such tournaments.

Table 7.4: Overview of activities

lesson	activity	concepts	representations
1	elephant estimation	average	picture, grid
2	life span of batteries	average, outliers, most, reliable	Minitool 1
3	compensation, mean strategy	mean accounting for all data values	value-bar graph
4	invention of battery brand data	mean (life span) and spread (reliability)	Minitool 1
5	estimating mean with median	mean and median, how are values distributed	row of numbers, value-bar graph
6	batteries, wing span, balloon	average, outliers average, sampling	Minitool 1, 2
7	growing samples in weight context	sampling, center, spread, shape	dot plot
8	speed sign (comparing distributions)	mean, median, majority	Minitool 2
9	reflection on distribution aspects	spread, sampling	value-bar graph, dot plot
10	jeans (single distribution)	sampling	several
11	jeans sizes	sampling, distribution	Minitool 2 (equal intervals)
12	final test	center, spread, sampling, shape	several

7.14.4 Summary of the answer

In short, the answer to the first research question is that students can learn to reason about distribution in an informal way in grade 7 when using an HLT similar to the one we used. The term 'distribution' must then be interpreted as "how data values are distributed" or as 'shape'. We conclude from the seventh-grade teaching experiments that the initial end goal "distribution as an entity-like object" was too ambitious. Students did not reason with distributions to solve statistical problems and did not operate with them as objects (except perhaps when reasoning about shifting bumps in class 1E). Most of the HLT's anticipations proved true, which is not really surprising after three other teaching experiments in similar classes. However, a few revisions still have to be made, especially around the median and the histogram activities. The results presented in this chapter answer not only the question of how, but also indicate to what extent *havo*-students can learn about distribution in about twelve lessons if conditions are favorable. In class 1B only a few episodes concerned reasoning with shape, but the level of reasoning did not meet the 1E episodes of Chapter 6, which are further analyzed in Chapter 8. In a future experiment we would spend more time on center and spread in relation to sampling, and less on the median, histogram, and shape. The activities of the first seven lessons (except the ones on the median) can be used to this end. The introduction of the median as a measure of center certainly requires much more time than we reserved for it. Using quartiles as a measure of spread requires even more effort, and probably has to wait until a higher grade. The same holds for introducing the histogram and box plot as ways to describe distributions. The growing samples appeared so viable for fostering coherent reasoning about the different key concepts that we decided to use it as a recurring theme in the eighth-grade experiment, which is the topic of Chapter 9.

8 Diagrammatic reasoning with the 'bump'

The real voyage of discovery consists not in seeking new landscapes, but in having new eyes.
Marcel Proust (1923/1929)[31]

The present chapter answers the second research question of how the process of symbolizing evolved when students learned to reason about distribution in grade 7. Because the end goal of the hypothetical learning trajectory (HLT) was reasoning about distribution in relation to other statistical notions and diagrams, we were especially interested in how students symbolized data into a 'bump', and how they reasoned with this bump as an object. What did the sign of a bump mean to students, and how did its meaning evolve in relation to this sign?
In the first section we establish why we did not use chains of signification (2.3.4), but turned to Peirce's semiotics to answer these questions (8.1). In 8.2 we define the Peircean notions that we need in the analyses of Section 8.3 and Chapter 9. Finally, we analyze the relevant classroom episodes and summarize an answer to the second research question (8.4).

8.1 From chains of signification to Peirce's semiotics

In Section 2.3.4 we have given our reasons for using semiotics to answer the research question on symbolizing. In short, semiotic theories study the process of meaning making in relation to signs. A sign, mostly something visible, signifies something else (signified, meaning, or object) that is mostly invisible. For example, people use a rose or a heart as standing for love, but also a sketch of the normal distribution as standing for that mathematical object. Signs are crucial in mathematics, statistics, and science, for instance, because learning these subjects and communicating about their invisible objects is impossible without signs. At the same time, and this makes mathematics and statistics so hard for students, the visible signs represent invisible mathematical objects or relations that students still need to learn.[32] Consequently, learning mathematics is a complex semiotic activity that requires both the construction of mathematical meaning and the interpretation and development of mathematical notation (Sáenz-Ludlow, 2003). Symbolizing is making signs that stand for those objects, but the objects also have to be formed (reification). The question we answer in this section is which semiotic theory best served our purpose of gaining insight into the symbolizing process when students learned to reason about bumps as objects.

31. This is not a literal quote but a summary of a much longer sentence.
32. This situation is referred to as the learning paradox (Bereiter, 1985). Hoffmann (2002) offers a solution to this paradox by using Peirce's semiotics.

As stated in Section 2.3.4, the semiotic framework we started with was that of chains of signification. The work of Cobb (1999) and Hall (2000) shows that chains of signification may be useful in analyzing relatively simple signification processes at the macro-level of an instructional sequence. Presmeg (2002) uses the notion of nested chaining for such signification processes. We initially used the chain of signification notion both as a design heuristic for the HLT and as an instrument of analysis. During the teaching experiments in grade 7, however, it turned out that this notion of a chain of signification was too linear in both the design and analysis phase. The following example illustrates this.

Figure 8.1: Emily's graph (left) and Mike's graph

After the second teaching experiment we tried to analyze the comparison of Mike and Emily's graphs (Figure 8.1) with the chain of signification theory. It was possible to reconstruct a chain leading to Mike's graph and one leading to Emily's graph, but the theory did not provide a solution to describe a comparison of the graphs with chains of signification. Yet the results of Section 6.11 show that it can be effective to let students compare different graphs,[33] for instance those of Minitool 1 and 2. In other words, we needed a theory that would be viable in network-like situations. We therefore underline Sfard's criticism of the metaphor of chains of signification as being too linear and simplistic for analyzing the reification process.[34]

> *Object*, therefore, is an aggregate of various optional attended and intended foci, brought into being by a collection of symbolic devices and discursive operations, organized experientially into one complex entity. Semioticians may be tempted to relate this entity to the idea of "chain of signification" (Walkerdine, 1988). Let me note, however, that the metaphor may be misleading inasmuch as it imposes linearity and thus oversimplifies the picture. (Sfard, 2000a, p. 322)

Searching for alternative semiotic theories, we first applied the system theory of epistemological triangles of Steinbring (1997). This theory addresses two aspects we considered important: the system character of signs and the network character of the learning process. By system character we mean that a sign, such as a dot plot, consists of sign elements (e.g. dots, axis, letters) and that this complex sign is also used

33. Cf. theories on multiple representations (e.g. Van Someren et al., 1998).
34. It is of course possible that people who use the metaphor of chains of signification do not interpret it as linear, but in our interpretation the metaphor of a chain is unfortunate.

in relation to other signs (e.g. bar graph, histogram, box plot). By the network-like character of learning, we mean that students interpret a sign system in relation to other systems and that they can go back and forth between different sign systems (Figure 8.2). In Steinbring's theory, a sign system is interpreted in the light of a reference system, and this can lead to the development of concepts (or mathematical objects). In contrast to the chains of signification theory, Steinbring's theory is non-linear and it allows for the comparison of different graphs.

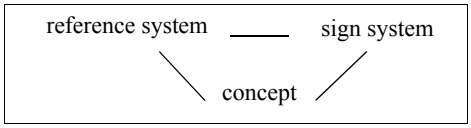


Figure 8.2: Steinbring's epistemological triangle

Looking back on the second teaching experiment and using the notion of reference systems, we conjectured that the learning process of several students was roughly what we informally represented in Figure 8.3. In Steinbring's terminology, Minitool 1 was meant to serve as a reference system for Minitool 2 (an arrow indicates 'being a reference system for'). Minitool 1 was a reference system for Emily's graph; Minitool 2 and line graphs were probably reference systems for Mike's graph. The comparison of those two graphs led to the notion of a bump and this bump was again used to interpret both Emily and Mike's graph as well as Minitool 1 (when students used the 'bump' to refer to the straight part in Minitool 1). This conjecture is analyzed more carefully in Section 8.3.

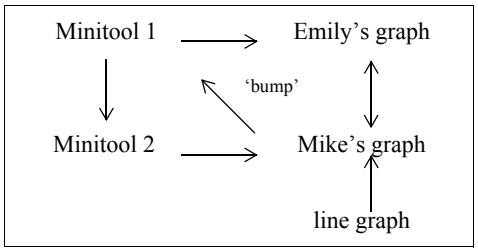


Figure 8.3: Network of sign systems in class 1E (an arrow means 'reference system for')

Although using this system theory highlights important aspects, it also raises a few theoretical problems. As elaborated by Hoffmann (2003b, in press), one of these problems is that it is not clear how the interaction between reference systems and sign systems leads to concept development; another problem is what the position of the knowing subject is in relation to the triangle.
All the theories[35] we have applied contributed to specific insights into the process

of symbolizing, but the theory that proved most insightful was the semiotics of Peirce.[36] The two aspects of his semiotics that lent themselves most to our purpose of analyzing students' learning were non-linearity and the possibility to stress the dynamic character of interpreting and making signs within the theory itself. We elaborate these aspects in the next section. There are also other attractive features. Peirce's semiotic framework offers a differentiated notion of sign and has a consistent epistemological basis. It has furthermore been developed in the context of mathematics, logic, philosophy, the history of science, and other disciplines he studied.
It is therefore not surprising that Peirce's semiotic notions apply more easily to mathematical signs and symbols than those of Saussure, whose sign notion has a textual origin (Ducrot & Todorov, 1983). Saussure defined language as "a system of signs" (Saussure, 1916/1974, p. 15), whereby signs for themselves were defined by a process of differentiating language as "a self-contained whole" (ibid., p. 9). Signs were thus defined exclusively by internal relations within language as a changeable system of signs. In contrast with how Saussure's sign notion is often used today (Mortensen & Robertsen, 1997), the famous distinction between 'signifier' and 'signified' referred *not* to a external representation and its referent: both concepts were defined by Saussure as purely psychological entities.
Though Peirce's semiotics is more easily applied to mathematical signs, it is by no means a ready-made instrument of analysis that can be applied to mathematics education. Peirce only slightly touched on education matters (Eisele, 1976). When applying his semiotics to understand the students' development of knowledge, we sometimes needed to stretch the original fields of application, but in doing so we stuck as closely as possible to Peirce's original definitions to avoid confusion and eclecticism. Sometimes we needed to choose from different definitions, because Peirce's views developed throughout his lifetime and he never wrote his 'final' epistemology (Ducrot & Todorov, 1983; Hoffmann, 2003a). Hence, Peirce's semiotics as presented below is a reconstruction for the specific purpose of analyzing students' learning.

8.2 Semiotic terminology of Peirce

In this section we elaborate on Peirce's semiotic theory insofar as we need it to answer the second research question on the symbolizing process. To explain the advantages of Peirce's sign notion as opposed to Saussure's, we need to define Peirce's sign notion and the different types of signs he distinguished. The key notions for an-

35. For instance, protocol and prototype (Dörfler, 2000), focal analysis (Sfard, 2000a, b), the transition from it, sign to natural object (Roth & McGinn, 1998; Roth & Bowen, 2001), and fusion (Nemirovsky & Monk, 2000).
36. We thank Michael Hoffmann for drawing our attention to Peirce's semiotics and for thoroughly commenting on Chapters 8 and 9.

alyzing the symbolizing process are 'diagrammatic reasoning' and 'hypostatic abstraction'.

Sign

Signs are at the heart of semiotics:

> All our thinking is performed on signs of some kind or other, either imagined or actually perceived. The best thinking, especially on mathematical subjects, is done by experimenting in the imagination upon a diagram or other *scheme*, and it facilitates the thought to have it before one's eyes. (Peirce, NEM I, p. 122)[37]

In Peirce's semiotics, a sign stands in a triadic relation to an object and an interpretant (CP 2.228). In contrast to the dyadic sign of Saussure consisting of signifier and signified, Peirce's sign involves an interpretant, which is the reaction or effect in acting, feeling, or thinking of the person who interprets the sign (hence the interpretant is not the interpreter). As Whitson (2003) observes, this reaction or effect need not be the necessary effect of a cause, but it is a sign-mediated response. This effect *can* be the production of a new sign.[38] The involvement of an interpretant makes it possible to highlight the idea of symbolizing as a dynamic activity.[39] This aspect is the first appealing aspect that we stress from Peirce's theory.

The second aspect is its non-linearity. An action or sign-mediated effect need not be a response to one single sign, but could be the response to several signs. Conversely, the effect of interpreting a sign can also be multiple actions or the production of multiple signs. Sign-activity therefore occurs within series, webs,[40] and networks of signs in which interpretants are responses to objects through the mediation of signs (Whitson, 2003).

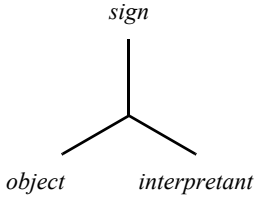


Figure 8.4: Peirce's sign in a triadic relation to object and interpretant

37. Following common practice, we refer to Peirce's Collected Papers as CP with the volume and section number, to the New Elements of Mathematics as NEM with the volume number, and to the Essential Peirce volumes of the Peirce Edition Project as EP.
38. This implies Peirce's theory allows a description of chain-like signification processes, but need not be confined to such linear processes.
39. Cobb (2002) and Walkerdine (1988) have solved this problem by focusing on the development of students' activities with signs.
40. Cf. the notion of webs of meaning (Noss & Hoyles, 1996; Salomon, 1998).

We give two examples to clarify how signs stand in a triadic relation to an object and an interpretant. If a student reads the sign '2 * 5 =' in an elementary school textbook, the interpretant can be the number 10. In that case, the interpretant is the result of calculating the sum. The interpretant is not a necessary effect as the student could make a mistake.
A more complex example is Whitson's (1997) umbrella example. Assume someone looks at a falling barometer (sign) and picks up his umbrella (interpretant). Presumably, the barometer reading is being interpreted as a sign of rain (object). Assume someone else sees him pick up his umbrella (sign) and also picks up her umbrella (interpretant). Others might decide, seeing the two leaving with umbrellas (sign), not to go out for lunch (interpretant). At a more detailed level, the barometer reading is already an interpretant which takes the needle position as a sign of atmospheric pressure. It is easy to extend this example to many other situations. For instance, the decision not to go out for lunch might be a response to the combination of seeing colleagues with umbrellas and listening to a weather report.
This umbrella example stresses that action is explicitly involved in Peirce's sign theory due to the interpretant (such as picking up an umbrella or the decision not to go out for lunch). This is why Peirce's notion of sign is sometimes characterized as more dynamic than Saussure's (Whitson, 1997). The example also illustrates the non-linear character of sign activity, because an interpretant can well be the response to different signs, and interpreting a sign can lead to different interpretants.
Peirce continued to reformulate his accounts until his death. For example, he was not consistent in defining a sign as one element of the triple (object, sign, interpretant) or as the whole triple. We chose to consistently use the first option of defining sign as one element of the triple because that is intuitively clearer to us, and it is closer to the everyday use of the term 'sign'.
There are two other aspects of signs we need to address. First, signs can be composed of other signs and they can be components of other signs (for instance, statistical diagrams are composed). Second, signs are used in different ways; that is, they can have different functions, depending on how they are interpreted as referring to objects. Peirce distinguished icons, indices, and symbols.

Icon

The key characteristic of an icon is its similarity to its object. Its main function is to represent relations. Icons represent things by imitation, for example photographs, but the resemblance may also be intellectual (Peirce, NEM III, p. 887). In CP 2.277, Peirce introduces three subcategories of icons: image, diagram, and metaphor (see also Stjernfelt, 2000). Diagrams are defined later in this section.

Index

The main function of indices is to direct someone's attention to something, exactly as in everyday language when we use the indices 'here', 'there', or 'now'. We can further think of a pointing finger and a thermometer, but also of demonstrative and relative pronouns, and of indices in algebraic formulas (such as the i in x_i) or in geometrical diagrams, such as the letters ABC to indicate the vertices of a triangle (Peirce, NEM III, p. 887).

Symbol

Symbols have become associated with their objects or meanings through usage, habit, or rule. Thus, if we interpret '5' as a sign for the mathematical object 5, it is a symbol.

> A Symbol is a sign which refers to the Object that it denotes by virtue of a law, usually an association of general ideas, which operates to cause the symbol to be interpreted as referring to that Object. (Peirce, EP, Vol. 2, p. 292)

Hence words and phrases are symbols as well as what is traditionally called a symbol in mathematics. The letter π, standing for the proportion of circumference and diameter of a circle, is a symbol; but note that the letter π printed on this page is not a symbol. Peirce used the terms 'token' and 'type' to make this distinction. An example he often used was the word 'the'. As a word, 'the' is a type (a symbol), but the instances on this page are only tokens of it.[41] For the analyses in this chapter we are especially interested in a particular sort of sign, diagram, because it is central to statistical reasoning as well as to mathematical and scientific reasoning.

Diagram

A diagram is a sign with indexical and symbolic elements, but its main function is iconic because it is used to represent relations. Peirce defined a diagram as follows.

> A *diagram* is a (sign) which is predominantly an icon of relations and is aided to be so by conventions. Indices are also more or less used. (CP 4.418)

Thus geometrical figures such as triangles can be diagrams because they represent particular relations of lines and vertices that are indicated by letters. Logical propositions are also diagrams, because they represent certain relations of other propositions, symbols and indices (e.g. the modus ponens: $(p \wedge (p \rightarrow q)) \rightarrow q$). Dörfler (2003) gives many other examples of diagrams in mathematics.

One reason why diagrams were so important to Peirce is that one can experiment

41. Token is therefore very similar to what many researchers today call an inscription. Furthermore, a similar distinction between token and type for geometrical diagrams is 'drawing' and 'figure' (Hoyles & Noss, 2003).

with them according to a certain syntax (logic, algebra, axioms, natural language, conventions of statistical diagrams), which can lead to new experiences (CP 5.9). This does not mean that all students have the same experiences, nor that they need to know all conventions and hidden rules of the diagrams they make.

Is a diagram a thing on paper or a computer screen, or is it a more general and symbolic sign? A diagram on paper is a token. If the relations of a diagram are interpreted as ideal, the diagram is a type. For example, if we prove that the angles of a triangle in Euclidean geometry sum up to 180^{o}, we use a geometrical diagram as a type, because we cannot prove any general or ideal relations from just the token of one particular drawing if it is not interpreted as standing for a triangle as a general mathematical object.

With this distinction we can clarify the close link between statistical diagrams and concepts that we mentioned in Section 2.2 and 2.3.4. If diagrams are just taught as tokens (how do you draw a box plot?) students are unlikely to conclude any general or aggregate information from them. Students need to develop concepts, otherwise they cannot reason with diagrams as types. This issue is elaborated upon under the headings of diagrammatic reasoning and hypostatic abstraction.

Diagrammatic reasoning

For Peirce, diagrammatic reasoning involved three steps.

1 The first step is to *construct* a diagram (or diagrams) by means of a representational system such as Euclidean geometry, but we can also think of diagrams in computer software or of an informal student sketch of statistical distribution. Such a construction of diagrams is supported by the need to represent the relations that students consider significant in a problem. This first step may be called 'diagrammatization'.

2 The second step of diagrammatic reasoning is to *experiment* with the diagram (or diagrams). Any experimenting with a diagram is executed within a representational system (not necessarily perfect) and is a rule or habit-driven activity (today we would stress that this activity is situated within a practice). What makes experimenting with diagrams important is the rationality immanent in them (Hoffmann, in press). The rules define the possible transformations and actions, but also the constraints of operations on diagrams. Statistical diagrams such as dot plots are also bound to certain rules: a dot has to be put above its value on the *x*-axis and this remains true even if for instance the scale is changed. Peirce stresses the importance of doing something when thinking or reasoning with diagrams:

> Thinking in general terms is not enough. It is necessary that something should be DONE. In geometry, subsidiary lines are drawn. In algebra, permissible transformations are made. Thereupon the faculty of observation is called into play. (CP 4.233)

In Minitool 2, for instance, students can do something with the data points such as organizing them into equal intervals or four equal groups.

3 The third step is to *observe* the results of experimenting. We refer to this as the reflection step. As Peirce wrote, the mathematician observing a diagram "puts before him an icon by the observation of which he detects relations between the parts of the diagram other than those which were used in its construction" (NEM III, p. 749). In this way he can "discover unnoticed and hidden relations among the parts" (CP 3.363; see also CP 1.383). The power of diagrammatic reasoning is that "we are continually bumping up against hard fact. We expected one thing, or passively took it for granted, and had the image of it in our minds, but experience forces that idea into the background, and compels us to think quite differently" (CP 1.324).

Diagrammatic reasoning, in particular the reflection step, is what can introduce the 'new'. New implications within a given representational system can be found, but possibly the need is felt to construct a new diagram that better serves its purpose (see Danny's example of symbolizing a distribution into three groups in Section 6.10). In 6.11 we saw how students' reflection on Emily and Mike's graphs led to something new, the bump, an abstract object that students referred to in different diagrams. This newly formed object can be viewed as an example of what Peirce called 'hypostatic abstraction', as we motivate in Section 8.3. This implies that anything that is made a clear topic of discussion or thought is an object. However, the object may be idiosyncratic (an 'immediate object') and need not be the culturally accepted concept (the 'final object').

Hypostatic abstraction

Peirce distinguished two types of abstraction, prescissive and hypostatic abstraction. Prescissive abstraction is dispensing with certain features; for example, if we use a geometrical line we dispense with the width of the line (CP 4.235). Hypostatic abstraction is making a new object.[42] For Peirce, "an 'object' means that which one speaks or thinks of" (NEM I, p. 124). A sign of hypostatic abstraction is that it puts "an abstract noun in place of a concrete predicate" (NEM IV, p. 160). This is not just a linguistic trick, but a genuinely creative act that allows someone to make discoveries (see this chapter's motto). Peirce considered this operation crucial in mathematics:

> (Hypostatic) abstraction is an essential part of almost every really helpful step in mathematics. (NEM IV, p. 160)

42. For Peirce, reification and hypostastic abstraction were the same; *hypostasis* is just the Greek equivalent of the Latin *reificatio,* the making of an object. If Sfard (1991) uses the term 'reification', it seems to apply only to the long-term process of object formation. Peirce's definition includes smaller steps.

Note that the term 'abstraction' can mean both the process and the product (cf. Noss & Hoyles, 1996, p. 123). If we want to stress the process of hypostatic abstraction, we do so by adding the word 'process'. Let us consider two examples to provide a clearer image of what this notion of hypostatic abstraction entails.

1 If we change "it is light here" into "there is light here" and consider 'light' as an object that we can talk about, we have a simple example of hypostatic abstraction (provided we only knew 'light' as a predicate and not as an object). In the first sentence 'light' is a predicate of something, but in the second sentence, 'light' is a noun, considered as an object in itself that can be predicated again. In the same way, we can transform the proposition "honey is sweet" into "honey possesses sweetness" (CP 4.235).
2 We give further examples of hypostatic abstraction from set theory. In the words of Peirce:

> In order to get an inkling – though a very slight one – of the importance of this operation (hypostatic abstraction) in mathematics, it will suffice to remember that a *collection* is an hypostatic abstraction, or *ens rationis*, that *multitude* is the hypostatic abstraction derived from a predicate of a collection, and that a *cardinal number* is an abstraction attached to a multitude. (CP 5.534)[43]

We interpret this as follows. Assume we have a batch of things, let us say, data values. We can conceive of these data values as belonging together or as being 'collected'. The notion of collection (or data set) is a hypostatic abstraction of the predicate 'belonging together'. We can further observe that there are many data values in this collection. A next hypostatic abstraction is to consider this predicate of many as an object that is interesting in itself and that can have its own properties: multitude. There are multitudes of different sizes (30 or 35 for instance) or, in other words, cardinal numbers. Cardinal numbers are hypostatic abstractions with their own properties, independent of the initial collection. As Sfard (1991) writes, the formation of such mathematical objects is by no means trivial to young children. And as we wrote in Chapter 5, students without a statistical background are not usually inclined to conceive of a data set as an object (hypostatic abstraction) that can have aggregate features. In our view, the most important goal in an instruction theory for statistics is that students come to develop hypostatic abstractions that can help them conceive of aggregate features of data sets. Forming a notion of distribution as an object (a hypostatic abstraction) can help students develop such an aggregate view. Another example of hypostatic abstraction that we discuss in Chapter 9 is that of forming a notion of spread. Students first use predicates to characterize the relative position of the dots in a dot plot as "the dots are spread out." The dots are signs that refer to data values (objects) and 'spread out' is a predicate of the dots. Later students write that "the spread is large." In that case, spread signifies a new object that is characterized with the predicate 'large'.

43. Note that such quotes were not meant in an educational but in a philosophical context.

The central point of diagrammatic reasoning is that it creates the opportunity for hypostatic abstraction: the formation of new objects such as those signified by the notions of dots, shape, or spread. Hypostatic abstraction can take place when a diagram (or part of it), or what it signifies becomes perceived as a new object that is signified by a new notion, such as a bump. This object can in turn be used as a tool for further operations in different contexts (for example, shifting bumps). These new objects are the starting points for further hypostatic abstractions, as the example on cardinal numbers shows.

Symbolizing

Now we have chosen another theoretical framework than we started with, we need to rethink what we mean by 'symbolizing'. In the literature on symbolizing (e.g. Cobb et al., 2000; Gravemeijer et al., 2002) the term 'symbolizing' is used for the process of making a symbol, using it, adjusting it and so on. The process of symbolizing and that of meaning development are assumed to co-evolve reflexively. Using the Peircean notions of diagrammatic reasoning and hypostatic abstraction we can describe how symbolizing in the Peircean sense evolves. Literally, symbolizing is 'making a symbol'. This includes both making a sign (such as a statistical diagram) and interpreting it as standing for an object (e.g. a distribution) in a conventional way. To symbolize, students therefore also need to develop a notion of that object (such as the bump) and interpret a sign as a symbol. This can happen during diagrammatic reasoning. In experimenting with diagrams and reflecting on the results, particular aspects of what is observed can become topics of discussion or thought; they can become objects by hypostatic abstraction in relation to the diagrams (or signs) at issue. If a diagram (or part of it) is interpreted as standing for that object, not in an iconic or indexical way, but by convention, it becomes a symbol. This means that symbolizing in the Peircean sense includes not only making signs but also the development of objects (for instance by hypostatic abstraction) and interpretants (how the signs are interpreted as standing for the objects). This means that the notions of diagrammatic reasoning and hypostatic abstraction provide tools for describing this process of symbolizing.

Before turning to the analysis of students' diagrammatic reasoning about the bump, we give a brief example of how we can analyze a symbolizing process by using the notions of diagrammatic reasoning and hypostatic abstraction.

Example of symbolizing

In Section 6.10 we described how a student, Danny, symbolized three groups of a hypothetical data set. The question was what a diagram of the class's height data would look like (this was before students had seen any data). Danny first made the first diagram in Figure 8.5 (this is an example of diagrammatization, the first step of

diagrammatic reasoning). When interviewing him, we initially interpreted his first sketch as a symbol for a normally distributed data set; in our view he even accounted for the variation around the smooth curve.

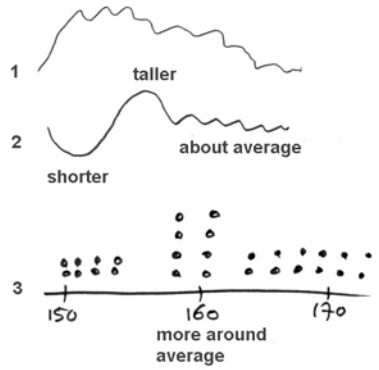


Figure 8.5: Symbolizing three groups of a hypothetical height data set

To test our own initial interpretation, we asked for clarification. However, he did not answer; instead his reaction was to make another sketch (which is the interpretant of our question). He explained this sketch with the words, “There are shorter, taller, and average students.” He probably imagined the students in row: first the shorter, then the taller, and last the average students. If so, he was mentally experimenting: imagining how the situation might appear. Trying to find out what he was thinking, we asked him where he had got this idea from. His reaction was again to draw another diagram, a dot plot. Again the interpretant to our question was a new sign. Pointing to the three groups of dots in his diagram he said, “There are short, tall, and average students; there are more around average.” This last aspect of his diagram is indicated by the higher stack of dots.

This is an example of diagrammatic reasoning because it includes diagrammatization, experimentation, and reflection. The hypostatic abstractions he formed or used during the mini-interview were the three groups he referred to: the group of short students, the group of tall students, and the group of average students. He used different ways to symbolize those groups (objects) in his second and third diagram. Doing so, he used conventions that he had learned during working with Minitool 2 and during class discussion in which there was reflection on the problems at issue. With this example we have shown how diagrammatic reasoning can lead to symbolizing, and which role hypostatic abstraction can play in this process. In the following section we analyze how students symbolized data into a bump by diagrammatic reasoning and by forming an abstract object that was signified by the bump notion.

8.3 Analysis of students' reasoning with the bump

Because reasoning about distribution as an object was the end goal of the research (except in class 1B), we focus the analysis on students' reasoning that came closest to this end goal: students' reasoning with the bump in class 1E (6.14). In this section we first analyze how the bump became a topic of discussion in the eleventh lesson. Then we look back at students' relevant experiences and look ahead to how students reasoned with bumps in later lessons.

The bump became a topic of discussion

Because it was Mike's diagram in Figure 8.6 that gave rise to the bump notion, we first analyze his diagrammatization, the first step of diagrammatic reasoning. Within the table, Mike interpreted data values ($signs_1$) as standing for students' weights ($objects_1$). The actions that he then undertook ($interpretants_1$) were to group the data values and count the frequency of each group. Mike then represented those groups of weight data values ($objects_2$) with dots ($signs_2$), and this led him to the next action of connecting the dots to an idiosyncratic line graph ($interpretant_2$). We write 'idiosyncratic' because it is an unconventional graph; for instance, the intervals are irregular.

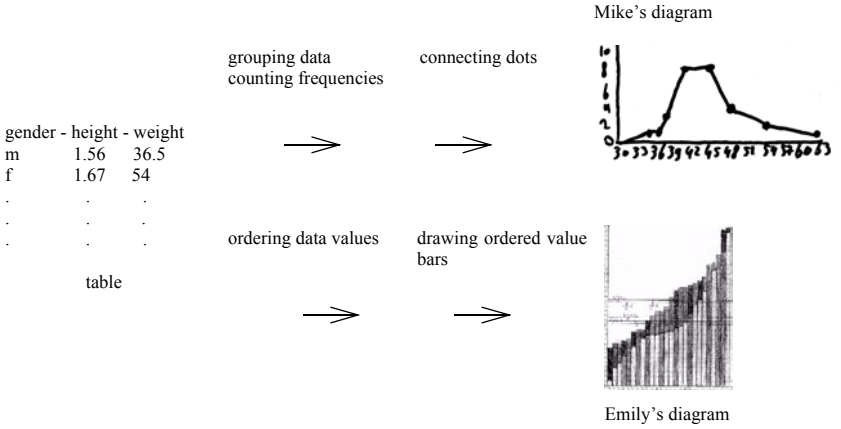


Figure 8.6: Reconstruction of Mike and Emily's diagrammatizations

In the HLT we had aimed for reasoning about shape assuming that it would support reasoning about the whole data set (an aggregate view) instead of just individual data values (a case-oriented view). When the teacher saw Mike's 'line graph', she realized that this was an opportunity to initiate a discussion on shape as intended in the HLT. Her reaction (interpretant) was that she used the term 'bump' ($sign_3$) to refer to the shape ($object_3$) in the graph. But what exactly is this shape? Depending on how it was interpreted, it could be anything ranging from a visual image to a symbol

of a slightly skewed unimodal distribution. In other words, the shape sign can have these different functions, as we clarify in the following.
The teacher herself probably interpreted the shape as standing for a unimodal distribution. The students probably first interpreted the sign 'bump' as a metaphor because of the resemblance of the shape with a bump. By asking what happened to the bump in Emily's graph, the teacher then stimulated students to reflect on the shape sign as a diagram representing relations between data values. For example, when the teacher asked about the bump in Emily's graph, Nathalie explained:

> Nathalie: The difference between … they stand from small to tall, so the bump, that is where the things, where the bars (from Emily's graph) are closest to one another.
>
> Teacher: What do you mean, where the bars are closest?
>
> Nathalie: The difference, the endpoints [of the bars], do not differ so much with the next one.

And Evelien added:

> Evelien: If you look well, then you see that almost in the middle, there it is straight almost and uh, [teacher points at the horizontal part in Emily's graph] yeah that [interrupted].
>
> Teacher: And that is what you [Nathalie] also said, uh, they are close together and here they are bunched up, as far as (…) weight is concerned.
>
> Evelien: And that is also that bump.

In our interpretation, the object that these students referred to was a group of values that were close together. The mental transformation of part of Emily's diagram into the bump of Mike's diagram or vice versa can be interpreted as a form of mental experimentation with diagrams. The episode therefore includes the three steps of diagrammatic reasoning: diagrammatization, experimentation, and reflection. This kind of diagrammatic reasoning offered an opportunity for hypostatic abstraction; in this case, a group of values that are closely together was referred to as a bump. This group is not yet a very definite object and it is very unlikely that students' interpretations of the bump were all the same. Yet the step of reasoning with bumps was an important step along the way to reasoning about distributions.

Where it came from
Mike's actions of grouping data values and drawing a line through dots have a history. In his explanation he talked about average, which might imply that his grouping action was inspired by averaging numbers in the first lessons of the teaching experiment. During the battery activity students had learned to talk about groups of data (e.g. "the high values of brand K"). It is also possible that his experience with the

Minitool 2 option of grouping data (e.g. making your own groups) inspired him to group the data. His next action was to make a *y*-axis with frequencies and to connect the dots as in a line graph (which, of course, is not a correct graph of the data). He had learned to make line graphs in mathematics lessons where such graph practices had been established. In the statistics lessons the students had never used frequency graphs before. As Walkerdine (1988) and Cobb (2002) note, signs such as the bump should be viewed within a particular practice, that is for a particular purpose with other mathematical practices in the background.

The background of Emily's diagram must be her experience with Minitool 1; she only turned the bars to a vertical position. The horizontal lines she drew represent the averages of boys, girls, and the whole class. This action stems from estimating averages with the value tool in Minitool 1 in earlier lessons. In other words, due to their experimenting with the Minitools, students were probably able to construct these diagrams and reason sensibly about them. In our view, the Minitools were useful for this experimentation step of diagrammatic reasoning because they offer the environment in which students can explore different options, organize data in different ways, and gain experience with the plots. In general, computer software might be especially useful during this experimentation phase, and perhaps also in the diagrammatization phase if the software has a user-friendly way of making diagrams.

When observing in the classroom, we thought that the bump referred to the distribution as we intended in the HLT. It was by going back to the history of students' actions and their exact formulations that we realized that students were just referring to the values that were close to one another and not to the whole distribution when using the term 'bump' (students also referred to this group as the 'majority' or the 'average'). The transcript lines above of Nathalie and Evelien, as well as transcripts of later lessons, support this claim.

The fact that several students in 1E were able to understand the link between Emily's and Mike's diagrams is probably due to their experience with the Minitools and the hypostatic abstractions they had formed before. By solving statistical problems with the Minitools, they had developed a language with which they could reason with[44] hypostatic abstractions such as majority, average, low and high values. It is likely that these notions formed the basis for interpreting the bump, because the bump was initially interpreted as the majority or average group of the data. Unlike the statistical convention, students called the data values of the remaining group 'outliers'. This shows that we had not had enough discussion on the correct meaning of some of the terms students used. The advantage of bump (or shape in general) over majority or average group is that it can also be used for the whole shape including 'outliers'.

44. We use the phrase 'reason *about'* if students mention properties of objects and 'reason *with'* if they use these objects as tools in their reasoning.

Where it went

In the twelfth lesson, when students revisited the battery problem, some students used the bump notion to indicate a specific group of the data and the straight part in Minitool 1. This indicates that they used the bump, a hypostatic abstraction, as a reasoning tool. For example:

Laura:	But then you see the bump here, let's say [Figure 8.7].
Yvonne:	This is the bump [pointing at the straight vertical part of the lower ten bars left of the letter D].
Researcher:	Where is that bump? Is it where you put that red line [the vertical value tool]?
Laura:	Yes, we used that value tool for it (...) to indicate it, indicate the bump. If you look at green [the upper ten], then you see that it lies further, the bump. So we think that green is better, because the bump is further up.

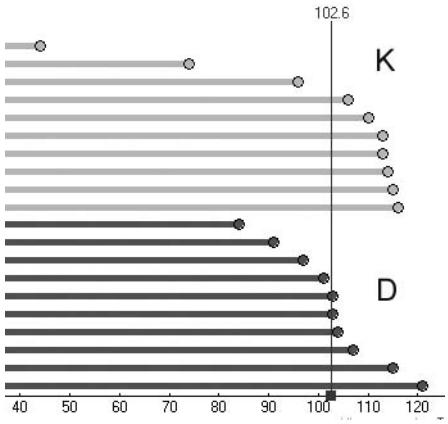


Figure 8.7: Reasoning with the 'bump' in Minitool 1.

These examples show that students did not interpret the 'bump' just as a metaphor because there is no similarity between the bump and the straight part in the diagram. The bump had become a symbol for several students, standing for the middle group of values that were close together, and this was by convention or habit grown out of the previous lesson. Apart from the term 'bump', students also used 'majority' or 'average' to refer to such groups of values that were close together. This 'majority' is not a very well defined object, but at least students talk and think about a group of data values with fuzzy borders.[45] Note, however, that it is possible to talk about the bump because there are also less frequent values that do not belong to the majority. They also used the term 'outliers' to refer to the remaining data values. The fact that

45. It is therefore useful to let students specify the range they regard as the 'bump' or 'majority'. Numbers and value tools can serve as tools for that.

Laura and Yvonne used the term 'bump' to indicate the straight part in Minitool 1 (Figure 8.7) shows that they had mentally constructed an object that was roughly the majority. In other words, they used the bump as a tool in comparing distributions.[46]
We now give more examples of students' experimenting with and reflecting on diagrams in the thirteenth lesson. In that lesson, the HLT aimed at letting students use bumps for the whole distribution and stimulating them to use shapes as tools in their reasoning. This aim was inspired by a remark of Sfard (1991) that to stimulate the formation of a concept as an object, we need to create a situation in which students need such a concept as an object and not merely as a procedure (or a batch of individual objects). We asked students about a larger sample to thematize the stability of the shape of a distribution, and we asked about the weight graph of older students to stimulate that students would shift the bump as an object (cf. Biehler, 2001).
In the thirteenth lesson, there were indications that several students came to relate the bump to the whole distribution instead of just the 'majority' or 'average' group. Emily, for example, incorporated the 'outliers' in her reasoning about the shape of the bump:

Researcher:	If you measured all seventh graders in the city instead of just your class, how would the graph change, or wouldn't it change?
Emily:	Then there would come a little more to the left and a little more to the right. Then the bump would become a little wider, I think. [She explained this using the term 'outliers'.]
Researcher:	Is there anybody who does not agree?
Mike:	Yes, if there are more children, then the average, so the most, that also becomes more. So the bump stays just the same.
Anissa:	I think that the number of children becomes more and that the bump stays the same.

Emily explained that the bump would become wider because of new 'outliers' (extreme values). This implies that she used the word 'bump' for the whole shape instead of just the majority or average group. What is interesting about Mike and Anissa's remarks is that they seemed to have a sense of the shape's stability even if the sample were to grow. If they had interpreted the bump as just the middle group, they would have probably thought that the bump might grow higher because of more dots.
It could well be that students started to use the term bump for the whole distribution because of the discourse on what students called 'outliers' (low and high values that

46. As an aside, once students have reached such understanding, they might come to see that the position of the bump can be measured with a median, because the median is generally somewhere in that bump, even in skewed distributions (unlike the mean). See also Cobb, McClain, & Gravemeijer (2003).

are less frequent than average values). The students could only see a 'majority' because there were values that occurred less often, which they called 'outliers'. Outliers are termed such because they differ considerably from the majority (the bump), and the bump is termed such because there are also values that do not occur so often. Mentally dividing data sets into three groups seems to be a natural step for students to think of data sets, as we inferred from the retrospective analyses (conjecture C1 in Chapters 6, 7, and 9).
The reasoning with bumps in larger samples could be seen as experimenting with a diagram in the mind. This also holds for students' reasoning in reaction to the following question, which was meant to stimulate a shift of the bump as an object.

Teacher: What would a graph of the weights of eighth graders look like? [instead of seventh graders]
Gerdien: The bump would be more to the right.

If Gerdien mentally shifted the whole bump to the right, the object she referred to was probably the whole distribution and not just the 'majority' of the data values. The question had created a need in which students could best operate with the bump as one object.
As we wrote in the previous section, diagrammatic reasoning offers the opportunity for hypostatic abstraction, the formation of objects (what can be talked about or thought of). In this section we encountered steps of hypostatic abstraction: the bump first as standing for a majority, and later for the whole distribution. In reasoning about the bumps students used hypostatic abstractions such as majority, outliers, and average, which they had formed in previous lessons, but which still needed refinement. This means that the process of developing a notion of distribution involves several steps of hypostatic abstraction and a gradual refinement of what the formed objects are. The way a few students reasoned about and with the bump shows that they developed an object that came close to distribution: they used the bump to model hypothetical data. Some probably realized that the shape would be stable across larger sample sizes.
Our intention was that students would finally come to say, for instance, "the distribution of brand D is normal" or "the distribution of brand K is skewed," or something synonymous. They did not. The closest to what was intended in the HLT was a dialog between two students (cited from Section 6.11.2) in relation to Figure 8.7 when they worked on one of the last tasks (revisiting the battery problem):

Anissa: Oh, that one is normal (…). [pointing in the diagram to brand D]
Nathalie: That hill.
Anissa: And skewed as if like here the hill is here [the straight part of brand K].

Anissa initially used a demonstrative pronoun to indicate something that is normal

or skewed; 'that one' hints at something but it is not clear what: is it the brand, the diagram, the majority, the shape, the distribution of the sample, or the distribution of the brand? We consider using words such as 'it' and 'that one' as a pre-stage to hypostatic abstraction, because 'it' is mostly used indexically in such cases. Nathalie interpreted 'that one' as a hill, a term which Anissa then used as well.
Though we regard the examples in this section as important accomplishments, they are too incidental to claim that the students of 1E had developed a notion of distribution as an object, as was intended in the HLT. In retrospect, we conclude that we had not been explicit enough on what we meant by "distribution as an entity-like object" (Chapter 5). One of the results of the research is in fact that we have a clearer image of the different levels of reasoning about shape and distributions.
At the start of instruction students already had a notion of what is normal and what is not, and they learned to express this in relation to different plots using terms such as outliers, average, majority, low and high values, et cetera. However, it was not until the thirteenth lesson in 1E that students used shape to refer to the whole data set. In Chapter 9 we provide examples in which eighth graders do reason about different kinds of shapes including 'normal', skewed, and bimodal.

8.4 Answer to the second research question

In this section we answer the second research question of how the process of symbolizing evolved when students learned to reason about shape in grade 7.
As mentioned in Chapter 7, students have a notion of what is normal and what is not. They can categorize familiar phenomena such as weight and height into low, average, and high values. During the teaching experiments, students worked with simple diagrams in the Minitools, value-bar graphs and dot plots, in which each data value is represented by one bar or dot. They experimented with these diagrams, solved statistical problems, reasoned about the data sets, and made their own diagrams (mostly similar to the Minitools). During this process of diagrammatic reasoning, students learned to describe aspects of the diagrams in terms of the situation (predication). For example, they talked about 'outliers', the majority, and average values. In retrospect, we concluded that we had not spent enough time discussing the meaning of some of these terms. For instance, students used the term 'outliers' for low or high values, not necessarily for suspect values or values outside the distribution, as statisticians would define outliers.
Students also learned to describe how data values were spread out. Battery brand K had more high values than brand D, but D was more reliable because the dots were less spread out. This implies that students formed several hypostatic abstractions during their diagrammatic reasoning in different contexts. Majority and average group match students' intuition of what is normal, whereas low and high values match what is not normal. The problem situations helped students interpret aspects

of the data represented in diagrams: if the dots were far apart, the brand was not reliable, and if the value bars were long, the brand had a long life span. Gradually, students learned to coordinate context issues, different representations (numbers, value-bar graphs, dot plots), and statistical notions. One way in which we stimulated this was by asking students to make diagrams of brands that had a long life span but were unreliable. We also asked them to make diagrams of large or small spread, and even to make a diagram with a large range but small spread.

We also argued that in comparing two representations students needed conceptual objects such as spread as tools in their reasoning, for instance when explaining which battery data set was which by looking at the way the data were spread out when comparing these data sets in Minitools 1 and 2 (see Section 7.3.2).

The analysis of the previous section shows that diagrammatization can involve multiple actions (interpretants). Mike, for example, informally grouped the weight data values, used dots at certain positions to signify these groups and their number, and connected the dots to one shape that was referred to as a bump. All of these actions have a history, either in the teaching experiment (e.g., grouping data or using dots) or in the mathematics lesson (e.g., line graph). The first interpretation of the bump might just have been a visual image or a metaphor, but the analysis shows that the meaning of the bump notion changed from the eleventh to the thirteenth lesson, at least for several students. In the eleventh lesson, the bump notion was used to refer to a group of values that were close together in the middle part of the graphs. The interesting thing is that the same data set looks quite different in various student graphs. By asking what had happened to the bump in Mike's graph in Emily's graph, the teacher stimulated students to formulate what exactly the object was which looked like a bump in Mike's graph and as a straight line in Emily's graph. This object, a group of values that were close together, can be seen as a hypostatic abstraction. Literally, the Greek *hypostasis* is 'what is underlying it' or the 'ground of things' (Muller & Thiel, 1986). We speak of it as the common conceptual structure underlying aspects of both graphs. If students had used aggregate plots such as histograms or box plots, it would have been unlikely that they could connect features of those plots to the group of individual data values that were close together. This supports the choice for relatively simple case-value plots such as value-bar graphs and dot plots.

In the twelfth lesson, several students used the term 'bump' even for a group of data if there was no visual bump, for instance when they referred to the straight part in Minitool 1-type representations as a bump. This implies that the bump was not just a visual characteristic, but had become a conceptual object and even a tool in their reasoning, for instance in arguing which battery brand was better.

In the thirteenth lesson, students referred to bump as the whole shape, whereas before they only referred to a group of values being close together. The development of the bump as an object was probably stimulated by what-if questions about hypo-

thetical situations in which students needed the bump as an object. When we asked about the shape of the graph with a much larger sample, one student argued that it would grow wider if the sample got bigger because there would be more outliers and other students reasoned that the bump would stay the same because there would also be more 'average' values.

Apart from the question of what would happen to the bump if the sample grew larger, students also answered the question of what would happen if students of a higher grade were measured. One student said that the bump would be shifted to the right. In that sense the bump had become an object encapsulating the data set as a whole. Several students were able to relate aspects of that shape to distribution aspects such as average, majority, groups of values, and several acknowledged the stability of the shape across sample size. They even hypothesized on the shape of a large sample, which means they developed a downward perspective of modeling hypothetical situations with a notion of distribution (5.2). Shape had become an object and reasoning tool for several students, but we cannot answer the question of whether distribution had become an object-like entity for the majority of students. To answer that question, we must specify what we mean by distribution, because there are multiple levels of understanding distribution. In a future teaching experiment, we could be more specific about which level of understanding distribution we will set as the end goal.

In retrospect, we can infer why it is so important to let students make their own diagrams and to let them explain them and reason with them. In constructing a diagram, students are likely to use implicit knowledge they have about diagrams, for instance where dots have to be placed. In explaining what they have done, they need to use words for the features of the diagrams they have created (groups of data, hill shape). Making such a feature (a predicate) a topic of discussion can lead to hypostatic abstraction, the formation of a more abstract object (e.g., majority, spread, or bump).

We have analyzed students' reasoning with the bump as a prototypical example of diagrammatic reasoning. Our experience is that several other episodes of students' reasoning with spread, range, Minitool options, and their own diagrams can also be analyzed with Peircean semiotics, so that the analyses contribute to our insight in the concept development.

Comparison with the Nashville answer

There are both similarities and differences between the results of the Nashville research and our own. In the Nashville research, the first "mathematical practice that emerged as the students used the first minitool can (...) be described as that of *exploring qualitative characteristics of collections of data points"* (Cobb, 2002, p. 179). Similarly, the students in the present study also started exploring qualitative characteristics of collections of data points with Minitool 1 (outliers, majority, aver-

age). The second "mathematical practice that had emerged as they developed these competencies can be described as that of *exploring qualitative characteristics of distributions"* (Cobb, 2002, p. 183). Some of the examples Cobb gives of this second practice are interpreting graphs of data sets that are organized into equal interval widths or into four equal groups, and that use reasoning in terms of global characteristics of distributions such as with a hill. Similarly, students in our teaching experiments came to reason with aggregate features of data sets and, in 1E and 1B, about bumps and hills. Cobb uses a chain of signification to describe how Minitool 1 served in the first mathematical practice as a signified for Minitool 2 in the second mathematical practice. We prefer to write that students formed hypostatic abstractions, such as average, majority, outliers, during the instructional activities with Minitool 1 for solving statistical problems with Minitool 2.

What is similar in our answer is that students indeed made progress from looking at data as individual data points towards reasoning with qualitative global characteristics of distributions such as with the bump in relation to average, low, and high values. What is different is the semiotic framework and the HLT that we used.

The semiotic framework that the Nashville team used, chains of signification, did not offer the possibility to compare representations and therefore did not suit our purpose of answering the question of how the symbolizing process evolved in a less linear HLT, in which students were stimulated to make their own diagrams and in which representations were compared. After applying several theories on key episodes, we ended up with a reconstruction of Peircean semiotics as an instrument of analysis that overcomes the linearity problem. In Chapter 9 we again use Peirce's semiotics to analyze how eighth graders came to reason about spread, shape, and distribution in a more advanced way than the seventh graders.

What can we learn from the analyses?

In this section we draw lessons from the retrospective analysis for the evolving instruction theory. We view these lessons as conjectures that can be tested in future teaching experiments. In short, students need to learn diagrammatic reasoning about distribution aspects. This implies several things for the three steps of diagrammatic reasoning, which need not happen in a particular order.

First, it is clear students need to diagrammatize–make their own diagrams that make sense to them, but also to learn powerful conventional diagrams. To stimulate aggregate views on data, we asked students to make diagrams according to aggregate features, for example of an unreliable battery brand with a long life span. This can be called diagrammatization according to aggregate features. The present study shows that students' own diagrams can be strongly influenced by the software to which they are accustomed.

Second, students need to experiment with diagrams. Educational software can be

useful in this stage of diagrammatic reasoning. The software should offer diagrams that students understand, but it should also offer opportunities for learning more advanced, culturally accepted diagrams. If the software is too directive or restrictive, students' creativity is constrained. If the software offers too many options, students might do a lot of things without making much sense of what they do. Then they might even take up undesirable habits that remain invisible to the teacher.

Experimentation need not only be done physically; it can also be done mentally. What-if questions can stimulate students to experiment mentally. Questions that can be asked include: What happens to the graph if a larger sample is taken? What would a graph of the class's height data look like?

Third, it is extremely important that reflection is stimulated, for instance by the teacher. Throughout the study we noticed that the best reasoning occurred during teacher-directed class discussions that were not in the computer lab, even though we tried to stimulate reflection in the instructional activities. Students are easily distracted by computers and while exploring data sets they are inclined to *do* things, not so much to *think* about what they do.

One of the core goals is that students learn to describe and predict aggregate features of data sets, because that is an essential characteristic of statistical data analysis. This implies that students should be stimulated to describe features of data sets and diagrams, and predict aggregate features of hypothetical situations.

Throughout this chapter we have shown that diagrammatic reasoning creates opportunities for developing concepts or, more generally, for developing hypostatic abstractions. This way of forming objects can be stimulated in different ways. First, predicates should become topics of discussion so that they can be taken as entities in themselves. For example, talking about 'most' data can lead to talking about the 'majority'; describing how dots are 'spread out' can lead to saying that the 'spread' is large. Second, students should be stimulated to be precise about what they refer to. For instance, If they use indexical words such as 'that' or 'it', it is possible that they cannot express or do not know to which object they exactly refer. In retrospect, we concluded we should have asked in which range exactly students saw the majority and where they saw the bump. Precisely defining the topic of discussion is thus integral to conceptual development. Third, we should create situations in which students need conceptual objects as reasoning tools (cf. Sfard, 1991). When the teacher asked what would happen to the diagram if data of an older class were shown, students were stimulated to use the bump as an object and shift it to the right as a whole. Fourth, comparing multiple representations (cf. Van Someren et al., 1998) of one data set can support students in thinking of the common structure (a hypostatic abstraction) underlying these representations. In the eleventh lesson, students compared several diagrams of one and the same data set and the teacher stimulated them to think of why the bump in one diagram looked different in another (Emily's). In explaining this, students referred to what the bump and the value bars stood for: val-

ues that were close to each other.

We do not want to suggest that these recommendations are easy to follow. In all of our teaching experiments it took considerable effort to promote a classroom culture in which students would participate well in discussions and would seriously try to write down their thoughts, and we often did not succeed. Nor does a practice of answering what-if questions emerge automatically. We noticed that when we visited teachers who used the seventh-grade instructional materials in the years after the teaching experiments we report on in Chapters 6 to 8. When we asked what-if questions or asked students to make diagrams of hypothetical situations during occasional visits, it was clear that they were not used to such questions and did not know what to do.

In Peirce's epistemologically based semiotics, diagrams are not only means of communication but, more fundamentally, means of thought, of understanding, and of reasoning. From this epistemological point of view, the essence of diagrammatic reasoning is that it offers the basis for hypostatic abstractions; cognitive means that can be used and developed in further diagrammatic reasoning (cf. Otte, 2003).

As we mentioned earlier, diagrammatic reasoning is not confined to statistics or mathematics. Modeling, an issue that receives much attention in different research communities, can also be framed as diagrammatic reasoning because a model is often a diagram as it is mostly used to represent relations. Modeling is then making a model, experimenting with it, and reflecting on the results. There are even attempts to use modeling or semiotic frameworks as alternatives to constructivistic frameworks (cf. Lesh & Doerr, 2003; Seeger, 1998). We therefore expect that the semiotic framework and type of analysis presented in this chapter have broad applications.

9 Diagrammatic reasoning about growing samples

I would like to compare this [process of symbolizing] with lignification [transformation into wood]. Where the tree lives and grows, it must be soft and sappy. If, however, the sappiness does not lignify, the tree cannot grow higher. If, on the contrary, all the green of the tree transforms into wood, the growing stops.

Frege in a letter to Hilbert (Frege, 1895/1976, p. 59; translation from German[47])

In Chapter 7, we answer the first research question for grade 7 and conjecture that students may develop a notion of distribution by reasoning about growing samples. Because distribution as an object-like entity proved too ambitious for grade 7, we conducted the next teaching experiment in grade 8. For the HLT of the teaching experiment in grade 8, analyzed in the present chapter, we decided to use the new activity of growing samples as the most important means of supporting reasoning about distribution. In Chapter 8, we answer the second research question for grade 7 and conclude that the notion of diagrammatic reasoning proves useful for analyzing students' symbolizing process when they learn to reason about distribution. It also appears possible to summarize the HLT for grade 8 as progressive diagrammatic reasoning about distribution aspects in relation to growing samples. Due to the HLT's formulation, the first and second research questions become strongly related. In the present chapter we therefore answer the two research questions simultaneously for grade 8 by answering the following integrated research question:

How can students with little statistical background develop a notion of distribution by diagrammatic reasoning about growing samples?

To answer this question we do not describe the whole teaching experiment, as in Chapter 7, but only the activities that were related to growing samples.
After providing information about the eighth-grade teaching experiment (9.1), we describe five episodes in which students reasoned about growing samples or shape (9.2 to 9.6). The results of the final interviews offer an image of students' notions of distribution after the teaching experiment (9.7). The last section summarizes the answer to the integrated research question (9.8).

9.1 Information about the teaching experiment in grade 8

In this section we supply relevant information about the class, the approach, the end goal of the HLT, and the method applied in the retrospective analysis in the eighth-grade teaching experiment.

47. Ich möchte dieses [Symbolisieren] mit dem Verholzungsvorgange vergleichen. Wo der Baum lebt und wächst, muss er weich und saftig sein. Wenn aber das Saftige nicht mit der Zeit verholzte, könnte keine bedeutende Höhe erreicht werden. Wenn dagegen alles Grüne verholzt ist, hört das Wachstum auf.

The class. The teaching experiment that we report on here was in an eighth-grade class with 30 students and lasted ten lessons of 50 minutes. After every second lesson we interviewed around eight students about their work (see also 3.7). The students had hardly learned any statistics except the mean and bar graphs, but during mathematics lessons students had learned about line graphs. The students were being prepared for pre-university (*vwo*) or higher vocational education (*havo*). In the Netherlands, the top 35-40% of the students attend *vwo* or *havo*. The remaining 60-65% are prepared for middle and lower vocational education (*vmbo*). In the practice of Dutch mathematics education, the school textbooks play a central role. Students are expected to work through the tasks by themselves, with the teacher available to help them if necessary. As a consequence, tasks are broken down into very small steps and real problem solving is rare. Students' answers tend to be superficial, partially because they have to deal with about eight contexts per lesson (Van den Boer, 2003). The students in the class reported on here were not accustomed to whole-class discussions, but rather to be "taken by the hand" as the teacher called it. The three assistants characterized the class as "passive but willing to cooperate." One of the challenges, therefore, was to get students engaged in statistical reasoning.

Guided reinvention. As mentioned in Section 2.1, we strove for a learning process of guided reinvention and for striking the balance between reinvention and guidance. On the one hand, we wanted to offer students opportunities to share their own ideas and participate in class discussions; and on the other hand, we wanted to guide their learning process towards culturally accepted statistical methods. This issue can be illustrated with a metaphor that Frege used in a letter to Hilbert (Frege was one of the first modern logicians and philosophers of language, and Hilbert was a formalist mathematician). The topic of the letter was using and making symbols in mathematical discourse.

> I would like to compare this with lignification [transformation into wood]. Where the tree lives and grows, it must be soft and sappy. If, however, the sappiness does not lignify, the tree cannot grow higher. If, on the contrary, all the green of the tree transforms into wood, the growing stops. (Frege, 1895/1976, p. 59)

Applied to our context of teaching statistical data analysis, this might imply the following. On the one hand, if statistical concepts such as mean, median, and mode are defined before students even have an intuitive idea of what these concepts are for, then just as the tree is at risk of transforming into wood, the students' conceptual development may be hindered. It could well be that students are not inclined to share their own ideas if they sense that statistical words have a very precise meaning that they do not yet understand. On the other hand, if teachers and instructional materials do not guide students well in a process of reinvention, the tree stays weak and cannot

grow higher. Students may even attach idiosyncratic meanings to statistical terms which are hard to 'unteach' again.

End goal of the HLT. As in the seventh-grade teaching experiment, the end goal of the HLT for grade 8 was that students would come to view distribution or shape as an object that can have different properties and that can be used to model data sets and statistical phenomena. As discussed in 2.2 and 5.2, distribution is an organizing conceptual structure with which students can conceive the data set as a whole instead of just individual data points. However, during the teaching experiments in grade 7, we had become more and more convinced that sampling should play a more central role than we had given it. We had noticed how important the process of talking through the process of data creation was as implicitly addressing the sampling issue. Attempts to make sampling more explicit, for instance with the trial of the Pyx and the balloon activities, had not really been very successful, but the growing samples activity in class 1B had proven promising. The goal of this eighth-grade teaching experiment was to test the conjecture that students could develop reasoning about distribution aspects including shape by reasoning about growing samples.
Compared to the seventh-grade experiments, we decided to do a few things a little different. We would pay more attention to the whole investigative cycle that is an important dimension of statistical thinking (Pfannkuch & Wild, in press): asking a question, design a study, take a sample, analyze data, and communicate the results. In the first lesson we let students first think of how one could test two battery brands. In several other lessons, we used newspaper reports and graphs to let students think about the way the information was acquired, which was also meant to address the sampling issue. For two lessons, students had to collect car color data to test a newspaper claim about percentages of car colors.[48] We also spent more time than in grade 7 on discussing the meaning of notions that had been the topic of previous lessons.

Retrospective analysis. For the retrospective analysis, the transcripts were read, the videotapes watched, and conjectures formulated on students' learning on the basis of the read and watched episodes. The generated conjectures were tested against the other episodes and the rest of the collected data (student work, field observations, and tests) in the next round of analysis (triangulation). Then the whole generating and testing process was repeated. Episodes in the transcripts that we considered examples or counterexamples of these conjectures were coded with these conjectures. The transcripts of lessons 4, 6, and 7 have been judged by three assistants. The amount of agreement was high: of the 35 code assignments that were discussed, only two were not judged unanimously. The conjectures that were confirmed are referred

48. In the Netherlands, in the fall of 2001, blue was the favorite car color with 21.9%, followed by red (20.8%), gray (18.9%), green (13.8%), and black (10.3%).

to as C# (see Appendix A). We give an example of an episode that was assigned with two conjecture codes. In the seventh lesson, two students used the four equal groups option in Minitool 2 for a revised version of the jeans activity (7.12).

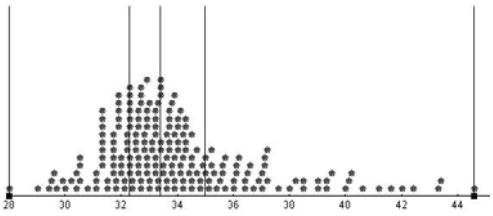


Figure 9.1: Minitool 2 with jeans data set organized into four equal groups

Int.:	Why did you choose four equal groups, Sofie?
Sofie.:	Because then you can best see the spread, how it is distributed.
Int.:	How it is distributed. And how do you see that here [in this graph]? What do you look at then? (...)
Sofie.:	Well, you can see that, for example, if you put a [vertical] line here, here a line, and here a line. Then you see here [two lines at the right] that there is a very large spread in that part, so to speak.

Sofie's answers were assigned with C7 and C2. C7 states that "students' notions of spread, distribution, and density are not yet distinguished. When explaining how data are spread out, they often describe the distribution or the density in some area." Related to this idea, C2 states that "students either characterize spread as range or look very locally at spread."
We also give an example of a code assignment that was dismissed in relation to the same diagram.

Int.:	What does this tell you? Four equal groups?
Melle:	Well, I think that most jeans are between 32 and 34 [inches].

We had originally assigned the code C1 to the this episode (students talk about data sets as consisting of three groups of low, 'average', and high values), because "most jeans are between 32 and 34" implies that below 32 and above 34 the frequencies are relatively low. In the episode, however, this student did not talk about three groups of low, average, and high values (or anything equivalent). We therefore removed the code from this episode.

9.2 Larger samples in the battery context

9.2.1 HLT for lessons 1-3

Sampling was an explicit issue from the first lesson onwards. For the first lesson, we decided to let students invent a method of testing two brands of batteries before any

data were provided. We would also stimulate students to think of aggregate features of samples larger than those they would think of: what would a sample[49] of a good brand look like? In the first lesson, the battery activity was used to stimulate discussion on features of both battery distributions (one normal and one skewed). To prevent students from being distracted by computers in the lab during a class discussion, we decided to have them switch off the monitors during discussions.
We did not use the elephant estimation activity as we assumed that these students, with more than one year experience in computing report grades, would already have a reasonably good sense of the mean. Moreover, we wanted to focus on the sampling conjecture in this teaching experiment and avoid the mean distractor effect (7.3.1).
The second lesson would be devoted to a discussion on students' answers on the battery problem. Just as in grade 7, the balloon problem would be used as preparation for the growing samples activity as used in grade 7.
In the third lesson, students would invent data sets according to specific aggregate features such as "brand A has a long life span but is not reliable." As in grade 7, this activity was used to stimulate students to represent aggregate distribution features (mainly average and spread) in graphical representations.

9.2.2 Retrospective analysis

One observation that struck us in the first lesson was that no student wanted to test more than two batteries per brand. Most thought one was enough; some added a second "for reserve." Even when prompted about this small number, many students thought that two batteries would be enough. Some were willing to test more if the machine in which they would be tested required more than two. We formulated the conjecture that students are inclined to think of small samples when first asked about how one could test something (C3). This was confirmed by the different data sources we had (audio, video, student work). A related conjecture that was confirmed (C4) was that students did not expect variation in this (industrial) context. This could explain why they did not think of larger samples.
From these observations we concluded two things. First, a notion of variability is a prerequisite for all statistical investigation (Chapter 5). Without a sense of variability, as in the battery context, there is no need of taking a sample or describing distribution aspects. Second, our approach of paying more attention to the design and sampling issue revealed that students did not expect variation in battery life spans of one type of battery. In grade 7, when we only talked through the process of data creation, this expectation had remained hidden. In hindsight, it would probably have been better to have chosen a context in which students do expect variation.

49. The Dutch term for sample is *steekproef*, which is a technical term most students of this age would not use in everyday language. In English the term 'sample' is informally used for examples of food or other products, but the Dutch language does not use the term *steekproef* for that.

Starting from students' ideas, we prompted them to think beyond the sample size of two. One question[50] that the interviewers asked was, "What if one brand had a good and a bad battery, and the other brand had two good batteries; what would you know about the brands?" From the analysis we concluded that such what-if questions proved useful to let students think about sampling in aggregate terms (C5). We give a few examples of students' answers to this what-if question.

Armin:	[Thinks a long time.] I would still think that the other brand [with two good ones] is better. But it could be coincidence. No, I don't know.
Int.:	What if you had measured 100 batteries and one of that first brand was bad?
Armin:	Then it could be coincidence, for example, but the fewer batteries you take, the more you doubt [*hoe meer je gaat twijfelen*].

Later he realized that it was impossible to test all batteries, and his peer said that it would be a lot of work to test many batteries. This shows that a short discussion can address sampling in a nutshell: sample should be large enough to draw reliable conclusions, but a large sample also has a price.
Another pair of students reasoned with proportions in answer to the same question, "What if one brand had a good and a bad battery, and the other brand had two good batteries; what would you know about the brands?"

Melle:	Then you still don't know very much.
Int.:	Why would you then take two [batteries]?
Sofie:	Yeah, I don't know, in the beginning we thought,... we had not really thought it through.
Int.:	What if you had measured ten and one was not very good. Would you know more?
Sofie:	Yes, you would actually.
Melle:	Yes.
Int.:	Or could it still be that one out of 100.
Melle:	In fact you cannot say anything, because [interrupted].
Sofie:	Actually, you could [*Eigenlijk wel ja*].
Int.:	When could you say something?
Melle:	If all [batteries] of one brand were good, and the other brand has seven bad and three good ones. Then you would be able to say something about it. Then the quality of the other one would be much better.[51]

When working with Minitool 1, students came up with arguments similar to those in grade 7 (such as: D has a higher mean, K has outliers, D is more together). There were also examples of case-oriented views, and certainly not all students had under-

50. This mini-interview question was not discussed before the lesson as we had not anticipated that students would all choose such small sample sizes.
51. On the basis of this episode and three others we conjectured that this kind of what-if questions could be used for making a connection with probability within a statistics unit. We have not yet been able to test this conjecture (cf. Biehler, 1994).

stood what the data stood for. It could be that some students did not think of samples because they did not expect variation in battery life spans (C4). Rick, for instance, wanted to buy the battery with the longest life span, the bottom bar in Figure 9.2.

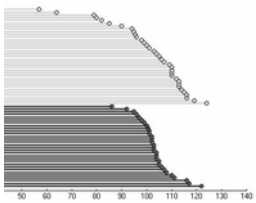


Figure 9.2: Battery data set in Minitool 1 (brand D: dark bars; brand K: light bars)

Int.: Rick, why would you buy the bottom one?
Rick: Because it goes furthest; it goes all the way beyond 120 [hours].

This underlines that understanding what the data stand for and describing them with aggregate terms was not self-evident for students. After a discussion on the reliability of their methods, students started to realize that it makes sense to take larger samples (they thought about ten would be reasonable).

In line with the what-if question of the mini-interviews in the first lesson, the teacher asked students a 'growing samples question' in the second lesson.

> Assume your two batteries lasted 85 and 105 hours and you took a larger sample. When is your conclusion 'the brand is bad' and when is it 'it is good'?

There were two types of answers. One type referred to spread only (example 1) and the other incorporated values or proportions of values (example 2). We give one example of each type.

> *Example 1.* The brand is good if all batteries live as long, if you take 10 batteries. The brand is bad if there is a lot of difference between the batteries, also with 10 batteries.
> *Example 2.* Good: if you test 10 batteries, there must be at least seven that last more than 100 hours and have about the same life span. Then there may be a few bad ones. Bad: if you test 10 batteries, and there are four bad ones, it is not a good brand.

Stimulating students with such what-if questions to discuss aggregate features in relation to samples had been reasonably successful if we take into account that this was the second lesson.

In the third lesson, students invented graphs in Minitool 1 of battery brands with specific aggregate features. Their inventions were similar to those in grade 7. We concluded that students were reasonably fluent in interpreting and producing distribution aspects such as average (life span) and spread (reliability) in this battery context. In terms of diagrammatic reasoning, students had diagrammatized specific aggregate features, mentally experimented with larger samples, and reflected on sample size in relation to features of battery brands.

9.3 Growing a sample in the weight context

9.3.1 HLT for lesson 4: towards shape as an object

The strategies for solving the balloon question that had been homework for the second lesson, had been similar to those in grade 7. Students had also made their first predictions of weight graphs. Thus the balloon activity prepared the growing samples activity in the fourth lesson. The overall goal of the growing samples activity in the weight context as formulated in the HLT was to let students develop a notion of distribution in relation to sampling. We conjectured that students could eventually conceive the stability of distribution shapes between samples as well as growing one sample, and that shape could become a topic of discussion. Our conjecture was that this transition from a discrete plurality of data values to a continuous entity of a distribution is important to foster a notion of distribution as an object. During teaching experiments in the seventh grade, in two American sixth-grade classes, and a visit to an American group of ninth graders, we observed that reasoning with continuous shapes turned out difficult to accomplish, even when we explicitly asked for it. It often seemed fruitless to nudge students towards drawing the general, continuous shape of data sets represented in dot plots. Our assumption was that students needed to construct something new, with which they can view the data or the phenomenon differently, for instance a notion of distribution (see the motto of Chapter 8).

Compared to the growing samples activity in grade 7, we would do a few things a little different. First, we decided not to let students measure their own weights because that could be too sensitive, but rather use real data sets from other classes. Second, to make sure that all students would formulate their own ideas (not just the ones that participated in a class discussion), we let them all write their comparisons down. We also tried to strike the balance between engaging students in statistical reasoning and allowing them to use their own terminology on the one hand, and guiding them in using conventional and more precise notions and graphical representations on the other.

9.3.2 Retrospective analysis

In retrospect we have come to see students' reasoning about growing samples as diagrammatic reasoning. As in Chapter 8, the process of hypostatic abstraction of shape appeared to consist of multiple steps. In this section, we analyze the lesson in three cycles, each consisting of making diagrams of a hypothetical situation (diagrammatization and a thought experiment) and comparing those sketches with diagrams displaying real data sets (reflection). In this section, we take all figures and written explanations from three students because their work gives an impression of the whole class's work in the following sense: their diagrams cover all types of diagrams made in this class and their learning abilities varied considerably. Ruud and Chris's report grades on the total of all subjects were in the bottom third of the class and Sandra had the best report grade of the class.

First cycle

The text of the activity sheet started as follows:

> Last week you made graphs for a balloon driver with data that you had invented yourselves. During this lesson you will get to see weight data of students from another school. We are going to investigate the influence of the sample size on the shape of the graph.
>
> a. Predict first a graph of ten data values, for example with the dots of Minitool 2.

The sample size of ten was chosen because the students found that size reasonable in the battery and balloon contexts. Figure 9.3 shows the different diagrams students made to show their predictions: there were three value-bar graphs (such as in Minitool 1), eight with only the endpoints (such as with 'hide bars'), and the remaining nineteen plots were dot plots (as in Minitool 2) (Fig. 9.4).

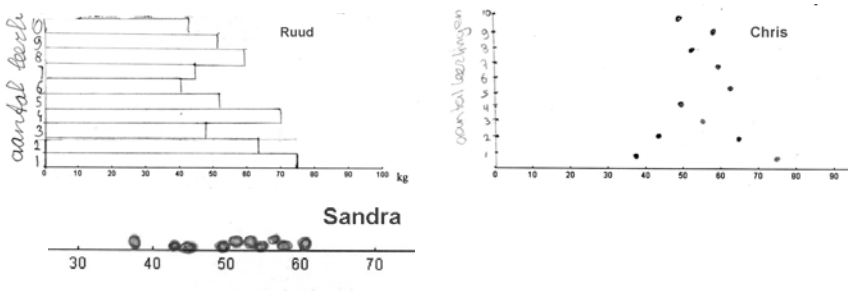


Figure 9.3: Student predictions for ten data points (weight in kg)

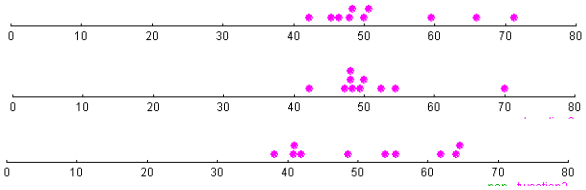


Figure 9.4: Three real samples in Minitool 2.

To stimulate the reflection on the diagrams (the last step of diagrammatic reasoning) the teacher showed three samples of ten data points on the blackboard and students had to compare their own diagrams with the diagrams of the real samples.

> b. You get to see three different samples of size 10. Are they different from your own prediction? Describe the differences.

The reason for the three small samples was to show the variation between samples.[52] We have found no clear indications, though, that students conceived this variation as a sign of sample sizes being too small to draw reliable conclusions, but they generally agreed that larger samples were more reliable.
There was a short class discussion on the diagrams with real data before students worked for themselves again. During this reflection activity, students were stimulated to describe features of the data sets, some of which were aggregate (together, further apart).

Jacob:	In the middle [graph] there are more together.
Teacher:	Here are many more together, clumped or so. Who can mention other differences?
Jacob:	Well, uh, the lowest, that is the furthest apart.
Teacher	Those are all furthest apart. Here they are in one clump. Are there any other things that you notice, Gigi?
Gigi:	Yes, the middle one has just one at 70.

The latter answer is a sign of a case-oriented view. The written answers of the three example students were the following.

Ruud:	Mine looks very much like what is on the blackboard.
Chris:	The middle-most [diagram on the blackboard] best resembles mine because the weights are close together and that is also the case in my graph. It lies between 35 and 75 [kg].
Sandra:	The other [real data] are more weights together and mine are further apart.

Ruud's answer is not very specific, as most of the written answers in the first cycle of growing samples were. Chris used the predicate 'close together' and added numbers that indicate the range, probably as an indication of spread.[53] Sandra used such terms as 'together' and 'further apart', which address spread. Many other students in this class also used daily-life words such as 'together', 'spread out', and 'further apart' to describe something that is both a property of the dots in the diagram and of the data.
This process of using predicates is also called predication[54] and can be considered a prerequisite to hypostatic abstraction: before 'spread' can be taken as a topic of common attention, a set of dots needs to be predicated with 'spread out.' For the analysis of the process of hypostatic abstraction it is important to note that the students used predicates (together, apart) and no nouns (spread) in this first cycle of growing sam-

52. See the research literature on resampling (e.g. Konold, 1994; Simon & Bruce, 1991).
53. Range was also historically the first sample measure of variability (David, 1998a).
54. Van Oers (2000) uses 'predication' in a slightly more specific sense: "Predication is the process of attaching extra quality to an object of common attention (such as a situation, topic or theme) and, by doing so, making it distinct from others" (p. 150).

ples. This changed in the second cycle of producing a diagram and comparing it with a real sample.

Second cycle

With the feedback of the samples of ten data points in dot plots, students had to make predictions for a whole class of 27 students and also for three classes with a total of 67 students (27 and 67 were the sample sizes of real data sets we had). During this cycle, all[55] students made dot plots, probably because the teacher had shown dot plots on the blackboard and because drawing so many value bars is laborious (Figures 9.4 and 9.5). In the seventh-grade experiment, we had left a lot of space for reinvention during the growing samples activity (7.9). In this case we wanted to guide the process a bit more, for instance by stimulating students to use statistical words. For research purposes, we also wanted to know what these terms meant to them.

c. We will now have a look how the graph changes with larger samples. Predict a sample of 27 students (one class) and of 67 students (three classes).

d. You now get to see real samples of those sizes. Describe the differences. You can use words such as majority, outliers, spread, average.

When the teacher showed the two real data sets (Figure 9.6), a short class discussion recurred in which the teacher capitalized on the question of why most of students' predictions now looked like what was on the blackboard, whereas earlier predictions varied more. No student had a reasonable explanation, which indicates that this was an advanced question.

The written answers to question (*d)* of the same three students were the following (Figure 9.5).

Ruud:	My spread is different.
Chris:	Mine resembles the sample, but I have more people around a certain weight and I do not really have outliers because I have 10 of about 70 and 80 and the real sample has only 6 around the 70 and 80.
Sandra:	With the 27 there are outliers and there is spread; with the 67 there are more together and more around the average.

In his written answer, Ruud addressed the issue of spread, although we cannot infer from this short answer what he meant by it. Chris was explicit about a particular area in her graph, the category of high values. Sandra used the term 'outliers' in this stage, by which students meant high or low values as we have seen in other classes. She also seemed to understand that many students are about average.

55. Only one student made a value-bar graph in the sample of 27, but she switched to a dot plot for the sample of 67.

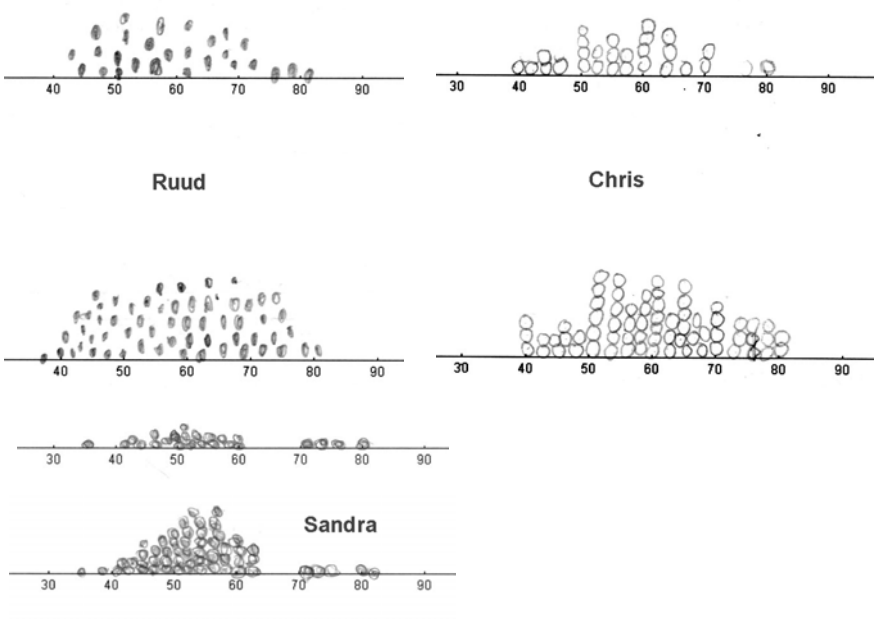


Figure 9.5: Predicted graphs for one and for three classes.

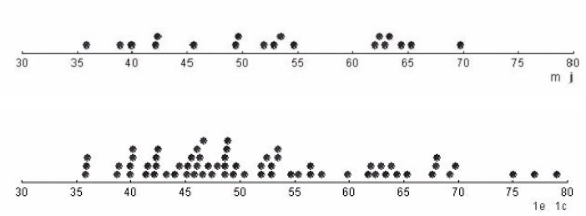


Figure 9.6: Real data sets of one and three classes in Minitool 2.

Sandra used the term ‘outliers’ in this stage, by which students meant high or low values as we have seen in other classes. She also seemed to understand that many students are about average.

These examples illustrate that students used statistical words to describe properties of the data and diagrams. From a statistical point of view, these terms were not very precise. By ‘mean’, students generally meant ‘about average’ or the ‘majority’; by ‘spread’, they meant “how far the data lie apart” (as in grade 7). By ‘sample’ they

seemed to mean just a bunch of people, not necessarily the data as being representative of a population (cf. Schwartz et al., 1998).
In contrast to the first cycle, students used nouns instead of just predicates to compare the diagrams. Ruud (like others) used the noun 'spread', whereas students earlier used only predicates such as 'spread out.' From a semiotic perspective this is an important step since it can be the sign of a hypostatic abstraction. This might seem a trivial linguistic trick, but statistically it makes a difference whether we say "the dots are spread out" or "the spread is large." In the latter case, spread is something that can have characteristics that can be predicated (large, small) or even measured (for instance, by the range or the interquartile range). Other notions, such as sample and average, were also used as nouns: that is, as objects that can be talked about. Recall that Peirce defined objects as things that could be talked or thought about. As the example of 'outliers' shows, the objects that students form during the process of hypostatic abstractions need not be the ones that we aimed for. From the context and from students' reasoning with notions as tools, we have to determine what they refer to when they use these notions.

Third cycle
So far, students had not talked about the shape of their graphs. In this last cycle of growing the sample, we asked for a graph that would show data of all students in the city, not necessarily with dots (Figure 9.7), and asked students to describe the shape of their graphs. The aim of asking this was to stimulate students to use continuous shapes and dynamically relate samples to populations without yet making that distinction explicit.

e. Make a weight graph of a sample of all eighth graders in Utrecht. You need not draw dots. It is the shape of the graph that is important.

f. Describe the shape of your graph and explain why you have drawn that shape.

The written answers to question f of the same three students were the following.

Ruud:	Because the average [values are] roughly between 50 and 60 kg.
Chris:	I think it is a pyramid shape. I have drawn my graph such because I found it easy to make and easy to read.
Sandra:	Because most are around the average and there are outliers at 30 and 80 [kg].

Ruud's answer resembles that of students in seventh grade who indicated a range of the average values or the majority. His answer focuses on the average group, or 'modal clump' as Konold and colleagues (2002) call such groups in the center. During an interview, Ruud literally called his graph a 'bell shape' though he had probably not encountered that term in a school situation before.

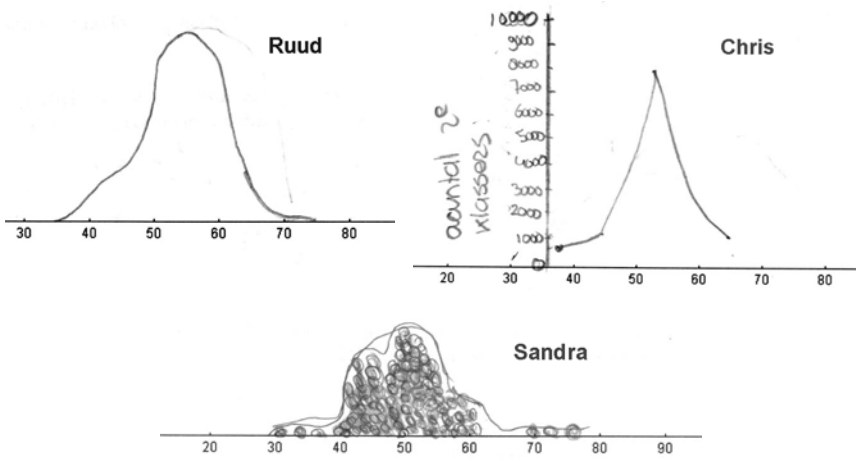


Figure 9.7: Predicted diagrams for all students in the city.

Chris's diagram was probably influenced by line graphs that they made during mathematics lessons (compare this with Mike's graph in Section 8.6). One of the aims in the HLT was indeed that students would draw continuous shapes of an informal notion of a probability distribution. However, this graph of Chris's shows the problem of using line graphs for that: line graphs cannot represent frequencies in the way Chris's diagram shows it because it is not clear what the class width is. Shape sketches without a vertical axis informally describe the course of the density, not the frequencies.
Sandra's diagram shows both dots and a continuous shape. It could well be that she started using dots and then drew the continuous shape. We had not anticipated such combined representations in the HLT. Her answer is an example of an episode that we coded with C1, on describing data sets as consisting of low, average, and high values.
In this third cycle of growing samples, 23 students drew a bump shape (mostly continuous). The words they used for the shapes were pyramid (three students), semicircle (one), and bell shape (four). Of course, we did not exactly know what these shapes meant to them. Therefore, students' reasoning with these shapes was taken up in the sixth lesson.

About the activity
This activity of growing samples involved short cycles of constructing diagrams of new hypothetical situations, comparing these with other diagrams of a real sample of the same size. We analyzed students' reasoning as an instance of diagrammatic

reasoning, which typically involves constructing diagrams, experimenting with them, and reflecting on the results of the thought experiments. Students' diagrams were strongly influenced by the two Minitools they had used (Minitool 1) and seen (Minitool 2), but they also used line graphs taught in mathematics lessons.

How did the process of hypostatic abstraction of spread evolve? Instead of just writing that the data were more spread out, students wrote or said that the spread was large. From the terms used in this fourth lesson, we conclude that many issues from Table 5.7, such as center (average, majority), spread (range and range of groups), and density (shape) had become topics of discussions (hypostatic abstractions) during the growing samples activity. Some of these words were used in a rather unconventional way, which implies that students need more guidance at this point. Shape became a topic of discussion as students predicted that the shape of the graph would be a semicircle, a pyramid, or a bell shape.

The growing samples activity combines different heuristics formulated in Chapter 6. First, it often stays away from data so as to avoid students from adopting a case-oriented view. Second, by asking students to compare their own diagrams with those representing real data, we invited them to "compare forests instead of trees", as the metaphor of another heuristic goes (Chapter 6). Third, by letting students predict a situation, we create the need to use conceptual tools for predicting that situation. The quick alternation between prediction and reflection during diagrammatic reasoning probably created ample opportunities for hypostatic abstraction.

In earlier lessons we had noticed that these students found it hard to concentrate during class discussions for longer than about ten minutes. A cycle of producing a diagram for a sample of a specific size, comparing it with a real sample requires short periods of concentration. Providing real data in between their inventions demanded short periods of reflection and feedback. We found it striking how well students knew the context of weight; their predictions resembled the actual samples in many respects. The delicacy of this subject might explain part of their engagement during class discussions.

9.4 Reasoning about shapes in the weight context

9.4.1 HLT for lesson 6: skewness as a topic of discussion

From the mini-interviews in previous lessons, we had concluded that students had a notion of distribution as consisting of three groups of low, average, and high values (C1). For example, Tula said, "A few low ones and a few high ones and more around average."

In the fourth lesson, almost all diagrams looked roughly symmetrical, which supports the hypothesis H21, which is based on the history of distributions, that students initially assume implicitly that distributions are symmetrical. In real life, however,

the phenomenon of weight shows distributions that are skewed to the right because of a "left wall effect" (two students had in fact drawn a left wall in the fourth lesson). By a left wall we mean that the lower limit (say about 30 kg) is relatively close to the average (53 kg) and the upper limit (sumo wrestlers can weigh up to 350 kg) is relatively far away from the average. This left wall in combination with no clear right wall causes the distribution to be skewed to the right. We therefore wanted skewness to become a topic of discussion as well.
In collaboration with the teacher, we invented the following activity to focus the students' attention to shape and skewness. We would draw the three student shapes on the blackboard and add two skewed shapes, which resulted in a pyramid, a semicircle, a bell shape, a unimodal distribution that was skewed to the right, and one that was skewed to the left (Figure 9.8). Students had to explain which shapes could *not* match the context of weight. We expected that it would be easier for students to engage in the discussion if they could argue which shapes were not correct instead of defending the shape they had chosen. Moreover, we anticipated a wider variety of reasoning than if all students defended one shape.

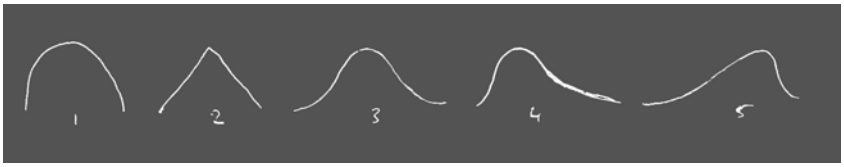


Figure 9.8: Five shapes as on the blackboard (1) a semicircle, (2) pyramid, (3) normal distribution, (4) distribution skewed to the right, (5) distribution skewed to the left. There were no axes with numbers.

9.4.2 Retrospective analysis

The teacher asked the class which shapes on the blackboard could not be right. For all shapes except the third, many students raised their hands. Apparently, most students expected a 'normal' shape (3). The teacher pointed out which students were to explain which shapes could not be right. Just as in class 1B we tried to involve all students in classroom discussions to avoid a situation in which just a small group revealed their thinking.

1 First, Gigi explained why the semicircle (1) could not be the right shape. One of his arguments was that there were too many people with low or high weights (the graph was relatively high at the endpoints of the shape).

Gigi: Well, I thought that it was a strange shape (…) I thought that the average was about here [a little more to the right than the top] and I found this one [peak of the hill] was a little too high. It has to be lower. And I thought that it was about 80, 90 [kg] here and I don't think that so many

	people weigh so much [points at the height of the graph at the part of the graph with higher values].
Teacher:	(…) Does everybody agree with what Gigi says?
Tom:	Yes, but I also had something else. That there are no outliers. That it is straight [vertical] and not that [he makes a horizontal gesture with two hands that looks like the tails of a normal distribution]. I would expect that it to slope more if it goes more to the outside [makes the same gesture].

Tom apparently understood that very low and high values do not occur very often, which means that the tails of the shape should be low and horizontal. Just as in grade 7, students used the term 'outliers' for low and high values that occur infrequently or for any values that deviated from the majority. As mentioned before, we concluded that we should have taken more effort to reserve that term for exceptional or suspect values and introduce another term for the tails of the shapes (such as low or high values; or the tails of the shapes).

2 Based on the students' reactions the teacher judged that they agreed that the semicircle was not the right shape, so she wiped it off the blackboard and turned to the pyramid shape (2). This discussion involved 'outliers' and the mean in relation to shape. It was Mourad's turn. So far students had only reacted to the teacher's questions–a very common type of classroom interaction (Van den Boer, 2003). In this lesson, however, students started to react to each other.

Mourad:	Well, I didn't think this was the one, because, yeah, I don't think that a graph can be that rectangular.
Teacher:	The graph is not so rectangular? [inviting him to say more]
Mourad:	No, there are no outliers or anything.
Alex:	It does have outliers; right at the end of both. It does have outliers.
Wim:	That's just the bottom [of the graph].
Alex:	At the end of the sloping line, there is an outlier, isn't there? (…)
Anna:	But the middle is the mean and everything else is an outlier. [Other students disagree, e.g. Fleur:]
Fleur:	Who says that the middle is the mean?
Anna:	Yes, yes, roughly then.
Teacher:	Tom, you want to respond.
Tom:	Look, if you have an outlier, then it has to go straight a bit [makes the same horizontal gesture as before]; otherwise it would not be an outlier (…) but that is not what I wanted to say. I wanted to react, that it [this graph] could not be the right one, because the peak is too sharp and then the mean would be too many of exactly the same.
Mike:	He just means that for exactly one weight all these kids weigh the same, so if the tip is at I-don't-know-how-many kilos, maybe 60 kilos, that all these kids are exactly 60 kilos.

In short, there were two aspects of the pyramid shape that students found inappropriate: the sharp peak (center) and the straight tails. In our interpretation,

these students understood that the frequencies of the weights around the mean would be similar; only if there were more of the exact same value as the mean, would the shape exhibit such a peak. We consider this kind of reasoning a form of mental experimentation in which students use statistical notions to reason what the cause of the sharp peak could have been. It is likely that their experience with dot plots in Minitool 2 and in lesson 4 enabled them to make this connection between the mean and the continuous shape. In retrospect, we often wished that we had asked students for more explanation: what do you mean by outliers? What do you mean by mean? However, it took quite some effort to have a class discussion and we decided to focus on the five shapes in the first place.[56]
Because the students agreed that this was not the right shape, the teacher also wiped this shape off the blackboard. As the interaction shows, students started to participate; their passive attitude started to change. Hearing the confusion between mean and mode, we decided to return to this issue at some point in the discussion.

3 Next, Sofie was to explain why the bell shape (3) could not be the right shape. Before the discussion almost all of the students thought this was the right shape (one girl admitted she did not know).

Sofie: I didn't have this as the one, because there are also overweight kids. Therefore, I thought that it should go a bit like this [draws the right part a little more to the right, thus indicating a distribution skewed to the right, Figure 9.5, shape 4].

The other students were not convinced. For instance:

Rick: That means that there are more heavier kids, but there are also kids who are underweight, so the other side should also go like that [this would imply a symmetrical graph].
Tom: Guys, this is the right graph!

Because several students still thought it was the right shape, the teacher did not wipe the normal shape off the blackboard.

56. Students were not always motivated to express their thoughts again and again: during interviews between lessons, a girl said, "In the beginning, I liked it, the first activities were pretty interesting, but then I thought that it was too much about one subject. Then it became a bit boring. Because, we had already discussed it, but you all kept asking about what we thought, though we had already explained it all. I found that a bit boring."

4 Next, Mike had to explain why the fourth graph could not be right.

Mike: I didn't think this was it because... if the average is, maybe, if this is the highest point, then this [part on the left] would be a little longer; then it would have a curve like there [left half of shape 3]. I don't think that this can be right at all, and I also find it strange that there are so many high outliers. Then you would maybe come to 120 kilos or so.

Some students argued that the mean need not be the value in the middle. Since students at this point argued about the mean versus the value that occurred the most, we decided to introduce a name for the mode, which these students had not learned before. Another reason we wanted to mention the mode (and in the ninth lesson the median) was that the instructional unit had to replace a schoolbook chapter on statistics in which the mean, median, and mode were addressed, which other teachers of the school were to teach. The teacher therefore needed to 'cover' these notions in this replacement unit. After the fourth shape had been discussed, we introduced a definition for the mode.

Researcher: The value that occurs the most often also has a name; it is called the mode [pointing at the value where the distribution has its peak]. (…) Who can explain in this graph [skewed to the right] whether the mean is higher or lower than the mode? (…)

Most students expected the mean to be higher (one raised her finger for lower). They found it hard to explain.

Tony: Most of what comes after it [the mode] is more than left of it, at the low side, so to speak. So there are more people with a high weight and few with low weight. [The students do not understand his explanation]
Res.: Who can say it again in his own words?
Rick: There are just more heavy people than light people, and therefore the mean is higher [in reference to shape 4].

Their remarks make sense if we interpret 'heavy' as heavier than the mode and 'light' as left of the mode and take into account that the right tail of shape 4 is further away from the mode than the left tail.
This is an example of how we used opportunities to introduce statistical terms, when students already talked about the corresponding concepts or informal precursors to them. Traditionally, the mode is introduced as the value that occurs the most in reference to data values (the upward perspective in Table 5.7), but we introduced it as a characteristic of a distribution (a downward perspective). Compared to the rather academic discussion of median and mean in class 1B, the discussion here of middle, mode, and mean was a heated debate in which many students became really engaged.

5 Last, Ellen said about the fifth graph:

Ellen: Well, I think this one is also wrong because there are more heavy people than light people. And I think that eighth graders are more around 50 kilos. That's it. [Note that there were no numbers in the sketch.]

Tom then objected that "nowhere does it say 50" and a lively discussion between the two evolved. We then asked Ellen to add numbers to her shape. She put 50 in the middle of the range, which would explain her saying that in this sketch "there are more heavy people than light people." Thus, as anticipated in the HLT, skewness became a topic of discussion in terms of heavy and light, even in relation to the mode and the mean, and the tails of the shapes, but not literally in terms of 'skewed.'

Students then wanted to know what the right shape was. Using a symmetry argument, we explained that it had to be something in between shapes 3 and 4.

Res.: Assume the shape were symmetrical, that left and right were exactly the same. Are there students in the Netherlands who weigh 90 kilos, do you think?
Students: Yes!
Res.: [Draws a shape with 50 as the mode and the right tail up to 90]. If it were symmetrical, then 10 kilos could also occur?
Students: No!
Res.: So it cannot really be symmetrical. It starts pretty low, 40 occurs. Perhaps even 30. And [there is] a tail to the right but not as long as in this shape [4].

About the activity

The aim of the HLT for this lesson was that students would learn to reason about skewed shapes, and they did so in terms of heavy and light. The satisfactory thing about this activity was that they came to reason with notions in a way they had not demonstrated before and that they were more engaged in the discussion than ever before, including the students with low grades for mathematics. We conjecture that the lack of formal rules, such as for manipulating algebraic formulas, makes it easier for low-achieving students to participate in the discussion. We furthermore conjecture that the lack of data, the game-like character and students' knowledge about the context were important factors, but also the fact that they had to argue against certain shapes. Such reasoning is safer than choosing the shape they think is right and defending that one. As we envisioned in Chapter 5, students started to develop a downward perspective on data: mode, average, and other statistical notions were discussed in relation to shape and were not just operations on data values.

With reference to the lignification metaphor (motto of this chapter), we could say that we had been successful in getting students to participate in reasoning about these

shapes. However, they often used terms (in particular 'outliers') in unconventional or vague ways. On the one hand, if statistical concepts are defined before students even have an intuitive idea of what these concepts are for, then students' conceptual development could be hindered (as discussed in Section 2.2). On the other hand, if teachers and instructional materials do not guide students well in a process of reinvention, the tree remains weak and cannot grow higher. It is evident that the notions of average, outliers, distribution, and sample of students in the present research needed to be developed into more precise notions, but at least students developed a language that was meaningful to them. They developed an image that could be sharpened later on; the sappy part of the tree that could be lignified in the future.
In terms of diagrammatic reasoning, this lesson was mainly devoted to reflection on shapes, but there were also examples of mental experimentation (what would the shape look like if...). Skewness was addressed within the weight context, but had not been predicated yet with terms such as 'left-skewed' or 'right-skewed.' Students mainly used two distribution aspects in their reasoning, average and the tails ('outliers'). These notions are hypostatic abstractions which have become reasoning tools. However, as we have mentioned before, students mostly had somewhat idiosyncratic but understandable interpretations of these terms. From the analysis we concluded that students probably had the following understanding of distribution: there are many values around average (high rounded part in the sketch) and few low and high values, which is evidenced by the horizontal tails of the shape.
In terms of emergent models (2.1), the shapes had become models of data sets such as weight. We envisioned that students would next use these shapes as models for a more mathematical reasoning about distribution, and would recognize shapes in other situations.

9.5 Growing the jeans data set in Minitool 2

At the end of the sixth lesson, we prepared the jeans activity (cf. 7.12) by asking students to design an experiment to find out about the waist sizes of men for a factory of jeans.
In grade 7, the jeans activity had turned out to be too complicated, partially due to the sampling issues involved. Nevertheless we expected eighth graders with more sampling experience to do better. Students were told they could earn 1500 guilders with a good report (this was just before the Euro was introduced). We made sample size an issue by offering students the opportunity to 'buy' samples. Small samples were cheaper than large samples. In the seventh lesson, students compared the waist data sets of different sample sizes in Minitool 2 (Figure 9.9). We cite from a mini-interview to illustrate how high-achieving students reasoned about the distribution aspects including shape.

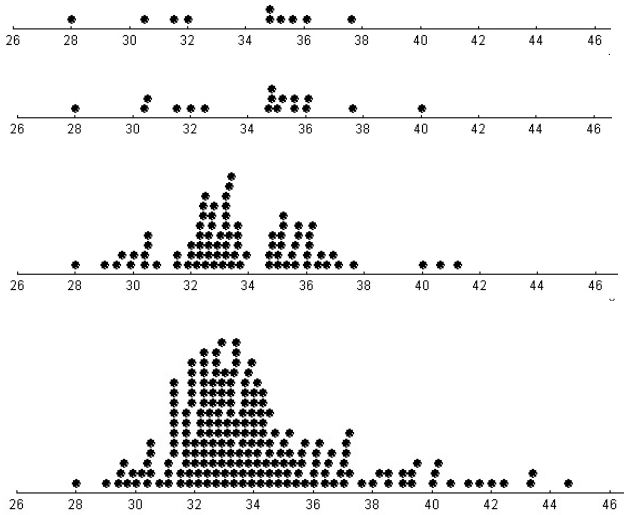


Figure 9.9: Jeans data sets of different sample sizes (10, 17, 82, 200); waist size is in inches. Students could scroll through them with the scroll wheel of the mouse.

Int.:	You had first opened the data set with 10 values. Can you tell me how it changed when you got more data [17]?
Tula:	Then you get a larger difference.
Int.:	What exactly do you mean by that?
Tula:	Well look. You can see that a bigger hump already merges here [around 32-34 inches].
Ellen:	And it is more spread out with more jeans because you have [inaudible].

Like most other students, these two mainly focused on two distribution aspects: spread (difference, spread out) and center (hump, majority). Despite the previous lesson, skewness (the position of the hill in relation to the other values) was not an issue that students addressed themselves.[57]

Int.:	What about the next one. What changed here [third graph with 82 dots]?
Tula:	Then you see that a real hill begins to take shape [*dat hier echt al een heuveltje begint te vormen*]. (...)
Int.:	And what happened to the spread? Is it larger or smaller?
Tula:	Well, there are more dots. (...)
Int.:	If you see this, what shape might you expect for a larger sample?
Tula:	Well, that there would be even more of a hill here.
Ellen:	And that something would come in between here as well.
Tula:	And here little less, so that's quite spread out.
Int.:	Ellen, why do you expect something to come in between here?

57. In one of the homework tasks students had to predict the shape of train delays, which is a very skewed distribution.

Ellen: If you have more people, then you'll also have people with that size.
Int.: Let's have a look. [They open the data set with 200 values.]
Ellen: Yeah, you see, I was right.

The mini-interview shows that Tula expected a hill to emerge from the growing sample and that Ellen expected the holes in the dot plots to be filled in. This example shows how two high-achieving students started to see shape as a pattern in the variability of the data. In hindsight, however, their reasoning did not appear to be representative of the whole class.

9.6 Growing samples from lists of numbers

9.6.1 HLT for lesson 8: recognizing shapes in dot plots

The HLT for the eighth lesson aimed at using shape as a tool in reasoning about distributions and sample size. The intention of the HLT for this eighth lesson was that students would learn that sample size is important and that the shape of a sample in a diagram will stabilize if the random sample is big enough. In the activity, students had to draw a growing sample from a set of 250 numbers, plot their data, and stop if they thought they knew the shape of the distribution (we had made uniform, normal, skewed distributions, but students did not know these terms yet). We expected students to recognize certain shapes. After this activity we showed the students dot plots of the 250 numbers of each distribution, and taught the terms 'uniform', 'normal', 'skewed to the left', and 'skewed to the right' in an informal way. We knew in advance that this activity without a context would not be easy for students, but we did not want to underestimate their abilities.

9.6.2 Retrospective analysis

This activity was not a success: students did not see the shapes in their samples and they mostly did not tell more than where the mode was. There were too many problems to ascertain what the core problem actually was. We mention a few obstacles.
First, the target shapes were often not visible from students' samples; they were not even apparent to us and we were privy to the source from which the samples were taken. Sometimes, the sample size was just too small, and sometimes students just had bad luck. For instance, Melle had taken 100 values from the first list and had gotten Figure 9.10, which looks far from uniform.
Second, there were also occasions in which we saw a right-skewed shape, whereas a student could not see it. For instance, Melle did not regard his fourth shape as skewed, although his peer Sofie did. Third, some students took an unsuitable scaling which resulted in no particular shape whatsoever. Fourth, there was no every-day context such as weight or height that could help students attribute meaning to the numbers. Fifth, in the sixth lesson students had reasoned about continuous shapes of population distributions, not about shapes of sample represented in dot plots. These

two situations were too different. In the latter case, there is a lot of noise around the signal of the shape.

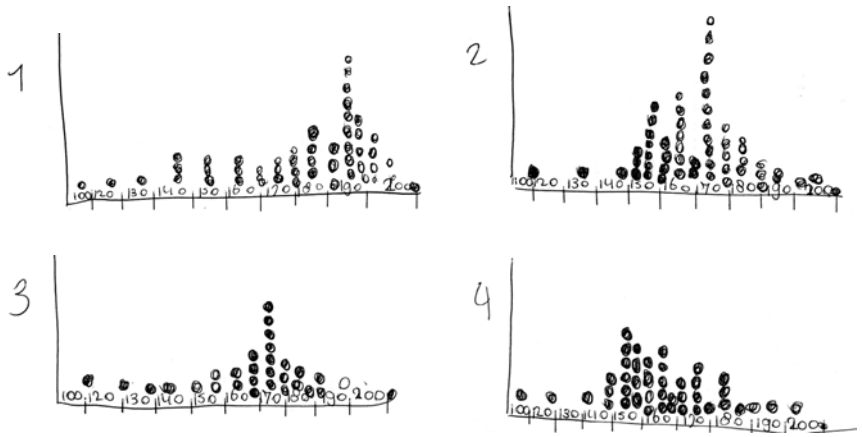


Figure 9.10: Melle's samples from the four lists (uniform, normal, left-skewed, and right-skewed).

In general, students did not interpret the diagrams the way we did, which indicates that they had not made the hypostatic abstraction steps we had expected them to make. Apparently you need to already have the shape in your mind to be able to see it in a diagram. Experts are inclined to separate signal and noise. The signal is the continuous shape with which they model the data, and the noise is the variation around the signal. Students had not yet learned to use the shapes they knew to model the samples they made, let alone to distinguish between signal and noise.

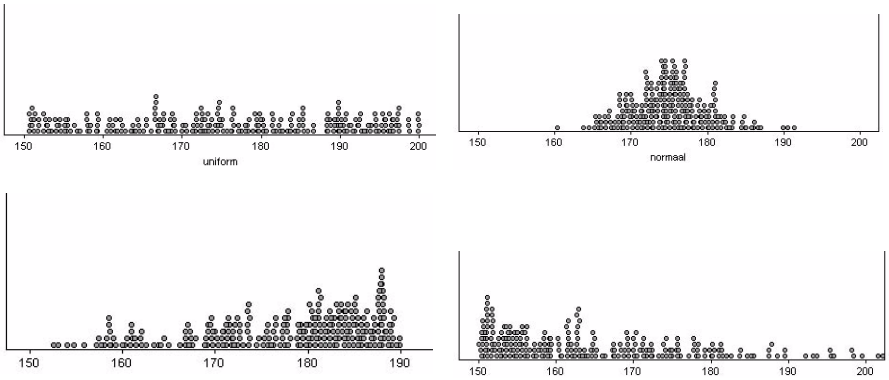


Figure 9.11: The populations of the four data sets from which students grew their samples (created in Fathom)

We mention these experiences as an illustration of a step in the HLT that was too big. The reasoning that students had demonstrated in the previous lessons appeared insufficient to recognize shapes in an abstract context. As in Section 8.4, the analyses of lessons 4, 6, and 8 also show that concept development is a gradual process with many instances of hypostatic abstraction. Additionally, the semiotic notion of diagrammatic reasoning, with experimenting as the second step, helped us to analyze what had gone wrong, as shown above. And from the fact students generally did not see the shapes that we saw in these diagrams, we conclude that they still had to perform certain steps of hypostatic abstraction. In semiotic terms, the same sign had different interpretants for the students than for us, because they had not constructed the same objects as we had in mind.
In Chapter 4 we mentioned the problematic relationship between phenomena and concepts as thought objects. We stated that people with different conceptual understandings can perceive different things. This section provided an example of that.
The analyses on diagrammatic reasoning about growing samples also indicate what the software could be useful for: experimenting with diagrams, the second step of diagrammatic reasoning. As shown in previous sections, students clearly made use of the two Minitool diagrams that they had used. In the present section it turned out that students had not had enough experience with an option that the software offers: scaling. If we had drawn more attention to this option and had highlighted it more in the HLT, the results may have been better. In terms of research, an advantage of this too large a step is that we were immediately able to see that these students certainly could not deal with much more advanced statistical problems.

9.7 Final interviews

In the final semi-structured interviews, we interviewed five pairs of students for about ten minutes per pair. In the fourth lesson, different shapes had been discussed and in the eighth lesson four types of distributions were given a name (uniform, normal, left-skewed, and right-skewed). Through these interviews, we wanted to find out what 'distribution' meant to students and whether they regarded measures of center as characteristics of a distribution. We decided to check whether students could correctly characterize a continuous sketch of a skewed distribution, and to ask them to indicate the position of mean, median, and mode. We discuss the results per question.

Table 9.1: Interview format

1	What is a distribution?
2	What kind of distributions are there?
3	Can you indicate mean, median, and mode in this distribution? (Figure 9.12) If students mistake the median for the midrange we also ask: Do you remember how you found the median in Minitool 2?

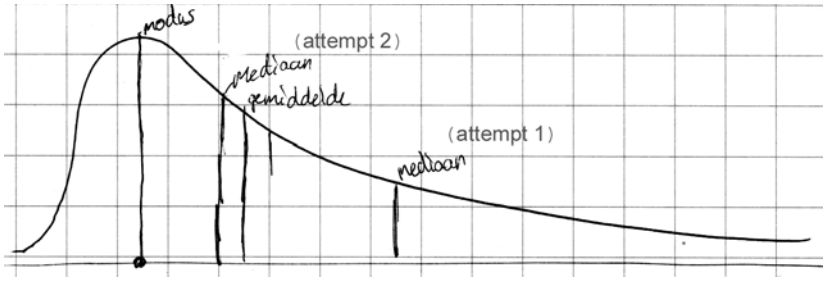


Figure 9.12: Sketch of a right-skewed distribution used for question 3. *Modus* = mode, *mediaan* = median, *gemiddelde* = mean. In the first attempt, the median was mistaken for the midrange.

What is a distribution?

Based on previous lessons we had expected that students would say something along the lines of "how the data are distributed in the graph." We present all the actual answers.

Pair 1. Steven thought that distribution was the "step size of the axis." Lynn added, "Then I think of the distribution of the data across the graph." Steven's answer probably refers to how the axis is labeled or how the data are structured, for instance in equal intervals. It could well be that he interpreted the Dutch term *verdeling*, which also has the connotation of division, as standing for how the axis was divided. Lynn's answer is an example of what we had expected.

Pair 2. Other students, such as Sanne, seemed to think of spread, or how the data are spread out. Her peer, Fleur, reacted with surprise.

Sanne:	If you have, for example, a graph with a lot of dots, then they can all be on one stack, so to speak. Then it is a very small distribution, because they all lie close to each other. But if they lie at the end and the beginning, then it is a very large distribution because they lie spread out across the line [the axis].
Fleur:	Is this the spread? Or is that the same? (...) I think the distribution is for example how you, how you for example uh, on this line, how you divided it, by 5, and then you get 10, 15, and then 20. How you divided it.

Where Sanne referred to spread, Fleur seemed to think of how the axis was divided (like Steven).

Pair 3. What Natasha and Sandra said came closer to what we had anticipated:

Natasha:	That is, uh, how do you say that. How, let's say everything (...). How everything, not the spread, but how the dots are in the graph.
Sandra:	How you put it then in the graph, the distribution of the uh, the dots then in the graph, where they are.

Pair 4. John described the process of getting a distribution, and it could be that he was referring to the transition from Minitool 1 to Minitool 2:

John:	Uh, … [laughs]. Well, you have a graph, with dots that have fallen down. And then at the bottom you have numbers, and the distribution is, how it is, how do you say that?
Int.:	Well, you already point it out, how it is...
John:	How it is put down, uh, for example, 20, 30, 40, 50, 60 and then the dots at uh, then you have for example 40 dots and there are for example fifteen of 20 and uh, fifteen of 60, and so on. But then you distribute it over it. That is the distribution.
Int.:	And what would you say, Ron?
Ron:	A distribution can also be straight, uh, be skewed, and uh, normal, that it is distributed evenly.

John's answer is interesting with respect to the process-product issue we address in Chapter 5. In that chapter we asked what the process aspect of distribution could be. Students' descriptions, such as John's might provide us with a clue: imagine that dots are distributed one-by-one over a variable; the result is the distribution. In fact, this is what students do if they grow a sample. We conjecture that the focus on growing samples in this teaching experiment has influenced student views on this process aspect of distribution or has provided them with a language to express their thoughts about distributions.

Ron was the only one who mentioned names of distributions learned the week before. Because students often used 'spread out' and 'distributed' interchangeably, we asked John:

Int.:	And what is spread again?
John:	You mean how the data is spread out or, in fact [laughs] or distributed over the graph.

This is another example of conjecture C7 that students' notions of spread, distribution, and density are intertwined. This supports the conjecture we make in Chapter 7: to develop a notion of distribution we might as well focus on spread and ask students to describe how data values are spread out, for example in dot plots.

Pair 5. Erno's answer also demonstrates this close connection of spread and distribution.

Erno:	Uh, distribution, that is uhm, pff, uh, how, yes, the data are spread out, are distributed.
Int.:	How would you say it, Alex?
Alex:	A distribution is the number of dots from 50 kilo to 100 kilo and then distributed [or divided], and that in the middle there is more for example.

Alex's answer hints at the image of distribution that we described in earlier lessons: a distribution as consisting of low and high values with low frequencies and a middle group with higher frequencies (C1).

What kind of distributions are there?

Most of the students did not remember precisely what the notions of uniform, normal, skewed to the left and right meant. All students except two called a distribution that is skewed to the right 'skewed to the left.' Most of the students who initially said 'skewed to the left' corrected themselves. Instead of uniform or normal they often used terms such as 'even' or 'equal' (*gelijk* in Dutch). Most of the time it was unclear whether they meant uniform or symmetrical. What is interesting from a semiotic point of view is that students generally do not say "the distribution is skewed to the right," but use demonstrative words such as 'that' or 'it' (which are indexical) as in "that one is skewed." In Section 8.3 we present an example of a student saying "that one is normal." Our impression was that students did not feel comfortable using the term 'shape' or 'distribution' itself. We consider this a pre-stage of a proper hypostatic abstraction.

Though it was interesting to see what students' definitions of distribution were, we think in retrospect that we should have stayed closer to what students had done in previous lessons. For instance, we could have asked students to describe the shapes in plots such as in Figure 9.11, to find out what they saw in them and how well they could describe aggregate features of those distributions. We could also have asked students to grow a sample from a large data set and to describe and predict aggregate features of the data set.

Measures of center in the skewed distribution

Students found it easy to find the mode, which is just where the distribution has its peak. However, they were not able to formulate it this way. For example:

Int.:	What was the mode again?
Fleur:	There where it is the highest. Isn't it?
Int.:	What is the highest?
Fleur:	Av... no not the average. How is it called?
Int.:	Do you know? [to her peer]
Sanne:	That line, where it has its highest point.
Int.:	Where the graph, the distribution has its peak.
Fleur:	That's what I meant.

Most could give a rough estimate of where the mean is. Their strategy seemed to be to start in the middle and look for high or low values that would influence the mean in one direction. For the median, all students except one pointed at the midrange, halfway between the minimum and maximum value. This happened often in the seventh grades (C8, see also Section 6.9). This is not really surprising because the median was only introduced on paper during the ninth lesson in a computer task, and was not further discussed by the teacher.

When students indicated the midrange as the median, we asked, "Do you remember

how you got the median in Minitool 2?" After such questions, several students concluded that there should be the same amount of dots left and right of the median.

> Fleur: So assume there are 100 dots in total, there must be 50 over there and then you put a line roughly there and then there are also 50 over here. [her estimate was quite accurate]

This could be called mental experimentation: Fleur made a mental connection between a dot plot and the continuous shape (see also Sandra's graph in Figure 9.7). Doing so she could accurately estimate the position of the median in the distribution. It is likely that students' experience with the two equal group option in Minitool 2 helped them to answer the question about the median. In our view, being able to locate mean, median, and mode in such a continuous sketch adds extra value to being able to calculate or determine these measures of center from a set of data. Located in a shape sketch, there is a meaningful relationship between these measures and the shape as a whole.

In short, students' notions of distribution could be characterized as follows. They had a sense of distribution as consisting of low and high values that occur infrequently and an average group with a higher density (we also discussed bimodal distributions in one of the lessons). They understand what the consequences of the differences in density are for the shape of the diagrams (low is infrequent and high is frequent). Students imagine a distribution as coming into existence through growing a sample (or collecting and plotting data): the dots are distributed over the axis with a specific pattern ("more in the middle," for instance). In this teaching experiment we tested the conjecture that students could develop a notion of distribution by diagrammatic reasoning about growing samples. In conclusion, we would say that this is indeed the case. By growing samples, students developed a sense of how distributions come into being, where hills emerge. There are indications that they see that holes get filled in at some point, and that the hill might become a little wider due to new extreme values.

9.8 Answer to the integrated research question

In the teaching experiment in grade 8, we tested the conjecture that students could develop a notion of distribution by reasoning about growing samples. From the results we concluded that this was indeed the case. This section summarizes an answer to the question of *how* eighth graders can develop a notion of distribution by diagrammatic reasoning about growing samples.

Distribution is a multifaceted concept that is difficult to learn, but learning to reason about distribution can be stimulated by the activity of growing samples as described in this chapter. We analyzed students' reasoning as instances of diagrammatic reasoning. What played a crucial role in this process is hypostatic abstraction as the for-

mation of objects. From the analyses presented above, it is clear that the development of the multifaceted concept of distribution included several steps of hypostatic abstraction.

We give a few examples of hypostatic abstraction from the previous sections. In the first and third lesson, students described what a larger sample of a good and a bad battery brand would look like. In this way, we stimulated predication and mental experimentation (a what-if attitude). In the fourth lesson, students used predicates such as 'spread out' and 'together' to say something about the dots (representing the data). Then they came to use nouns such as 'spread', 'average', and 'outliers', which indicated steps of hypostatic abstraction of distribution aspects, and they used these nouns as tools in their reasoning about shapes of distributions. Students started to see shape as something that they could talk about, which indicates another instance of hypostatic abstraction. Students guessed about the shape of the weight distribution and came up with semicircle, pyramid, and bell shape. It must be due to their experimenting with value-bar graphs (in Minitool 1) and dot plots (Minitool 2) that students were able to construct the diagrams of Figures 9.3 to 9.7 and reflect on them.

In the sixth lesson, students discussed these shapes and used statistical notions such as mean and outlier to explain why certain shapes could not represent a weight distribution. During the diagrammatic reasoning about this, the statistical notions of mean and 'outlier' were used as reasoning tools. Students also implicitly reasoned about frequency and density in this phase. To make skewness a topic of discussion as well, we introduced two skewed shapes.

In the eighth lesson we discovered that we had made too big a step in the HLT. Although students had learned to reason about continuous shapes in lesson 6, they found it difficult to recognize shapes in growing samples represented in dot plots without a meaningful context. If we semiotically compare how students interpret the shapes in lessons 6 and 8, it becomes clear that students had not yet made certain hypostatic abstraction steps necessary to recognize shapes in dot plots with a lot of variation instead of the continuous shapes used in lesson 6. Additionally, what had been missing was a specific type of experimentation with diagrams: scaling. Without a proper scale it is hard to see the shapes of the distributions. Yet students learned to describe shapes and distributions as being uniform, normal, skewed to the left or to the right. This means that shape had become an object they could reason about.

One of the end goals of the HLT was that distribution would become an object-like entity. In Chapter 5, we argue that a distribution is more like a composite unit than an object with a procedural and structural side and wonder what the procedural side of a distribution would be. From the way students talked about distribution, in particular during the final interviews, we inferred that they imagined the process of distributing dots over the variable as if growing a sample in a dot plot. We conjecture that such a process view of distribution could well be the procedural side of the concept, but we realize that our focus on growing samples has also fostered this view of

distribution.

The analysis shows that the reification process of distribution is a complex process that involves many steps of hypostatic abstractions. Understanding distribution requires understanding key aspects such as center, spread, density, and skewness. There even seems to be a reflexive relationship between the development of such characteristics of a distribution and the notion of distribution as an object or a shape: by reasoning about the occurrence of low, average, and high values, students expect a particular shape, and by reasoning about shape, students develop the meaning of distribution aspects such as mean, spread, density, and skewness.

We cannot answer the question of whether distribution had indeed become an object for the majority of the students without specifying what we mean by distribution and by object. In Section 10.1.3 we specify different levels.

As a final remark we would like to stress that, for instructional design to be successful, it is not enough that the instructional materials are well designed. The fact that the growing samples activities turned out successful was, in our view, due to a balance between students' background in mathematics, the Minitools, the teacher's ability to orchestrate discussions, and the timing of the activity. In line with findings of Kanselaar, Van Galen, Beemer, Erkens, and Gravemeijer (1999), we contend that these variables cannot and should not be investigated as separate factors. The methodology of design research offers a way to investigate these issues coherently. The metaphor that came to mind is the sound of a symphony orchestra. One oboe player playing too high, a trumpet player playing too loud, or a timpanist playing too early can ruin the chord as a whole. A chord only sounds good if all musical aspects are in tune with each other. The advantage of design research is that this tuning process can be accomplished in different cycles of anticipation and adjustment. As a consequence, we recommend to invest in using computer tools only if other factors such as teacher support, instructional activities, end goals, and assessment are adjusted to using these computer tools.

10 Conclusions and discussion

Statistics is a little arithmetic and a lot of thinking
Eighth-grade student

The purpose of the present research is to contribute to an empirically grounded instruction theory for early statistics education. An instruction theory, in short, is a theory of how students can be supported in learning a specific topic, in our case the concept of distribution in relation to other statistical key concepts and graphical representations. The contribution of the present study to such an instruction theory is summarized in this final chapter, which consists of a reflective and a prospective component. In the reflective component, we summarize the answers to the research questions (10.1) and other more general results relevant to an instruction theory (10.2). In the discussion of the results, several topics are addressed: methodology, heuristics for Realistic Mathematics Education, computer tools, and symbolizing; and a comparison is made with the Nashville research (10.3). In the prospective component, we make suggestions for a statistics curriculum at the middle school level (10.4). The last section offers recommendations for teaching, instructional design, and future research (10.5).

10.1 Answers to the research questions

The research questions of the present study are:

1. *How can students with little statistical background develop a notion of distribution?*

2. *How does the process of symbolizing evolve when students learn to reason about distribution?*

The first question is answered by summarizing a reconstruction of the hypothetical learning trajectory (HLT) on the basis of what has been learned from the study. This implies an omission of the activities that were not so fruitful within the trajectory (Section 10.1.1). As shown in Chapters 8 and 9, the process of symbolizing can be understood as embedded in the process of diagrammatic reasoning. The second research question is therefore answered by summarizing the key steps in students' diagrammatic reasoning about distribution aspects, and about shape in particular (Section 10.1.2). During the teaching experiments in grade 7, the idea emerged of using the activity of 'growing samples' to support students' development of the notion of distribution. In retrospect, it proved possible to frame the HLT for grade 8 as progressive diagrammatic reasoning about distribution aspects in relation to growing

samples. Due to this formulation, the first and second research questions became strongly related. In grade 8, we therefore answer the two research questions simultaneously by answering the following integrated research question:

How can students with little statistical background develop a notion of distribution by diagrammatic reasoning about growing samples?

The answer to this question is summarized in Section 10.1.3.
As a background to the answers, we summarize relevant information about the design of the HLTs. The design of the first HLT was prepared by a historical and a didactical phenomenology (Chapters 4 and 5). The historical phenomenology was carried out to gain ideas for instructional activities and to formulate hypotheses about students' statistical learning. One of the hypotheses was that estimation tasks could support an implicit use of the mean. As part of the didactical phenomenology, we then translated the historical contexts into modern ones so as to be useful for education. In the didactical phenomenology, we identified the core goal of the HLT as extending students' case-oriented views with aggregate views of data. Building on ideas of the Nashville team (Cobb, McClain, & Gravemeijer, 2003), we assumed that a notion of distribution ('shape') could offer a conceptual structure with which students could come to reason about aggregate features of data sets.
The one-sentence summary of the HLT was to challenge students to reason about distribution aspects with increasingly sophisticated notions and diagrams. The initial ideas for the HLT were elaborated in several teaching experiments in grade 7 (Chapter 6) and tested in the last seventh-grade teaching experiment (class 1B; Chapter 7). In the HLT for 1B we initially focused on center and spread as the most important distribution aspects and then factored everything together by discussing growing samples and by making shape a topic of discussion. In class 1B, students came to characterize shapes as bumps and hills, but their reasoning was not as sophisticated as in class 1E (Chapter 8). Because one instructional activity, called growing samples, turned out to be particularly fruitful to foster coherent reasoning about distribution aspects, we decided to take this activity as the recurring theme in grade 8 (Chapter 9).

10.1.1 Answer to the first research question

This section provides a reconstruction of the HLT tested in the last teaching experiment in grade 7. For a schematic overview of the HLT reconstruction see Figure 10.1. The claims have to be taken as anticipatory conjectures that can be tested and revised in practice (cf. Section 3.1).

1. Estimate a large number or objects in a picture. When estimating the size of a

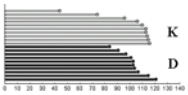
herd of elephants for instance, students need to find a way to use part of the picture to find the total number. One of the possible strategies is to make a grid and count the elephants in an 'average box', which is the box in the grid representing the other boxes such that the total is accurate. In discussing what an average box is, students learn to look at how values are distributed. The average box is also a sample that is chosen because its size helps to say something about the total. The issue of representativeness can thus be dealt with in an intuitively clear manner. It is possible that students use a midrange strategy of averaging the fullest and emptiest box in the grid. Skewed distributions in other grids can then be used to challenge the midrange as a measure of center and focus students' attention to how values are distributed in relation to the center.

2. Compare different distribution aspects in a value-bar graph. To avoid the so-called 'mean distractor effect', distribution aspects other than the mean have to be addressed (cf. Kelly & Beamer, 1986). The two battery data sets were chosen based on similar means, but different distributions: one is normal and one skewed. In this way, students are stimulated to describe other features of the data sets than the mean. They are likely to reason about extreme values, values that are close together, and the reliability of the brands. In other words, they start to reason about spread issues.

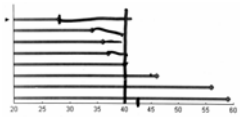
3. Explain what an average box is in a value-bar graph and reinvent a compensation strategy for visually estimating the mean. The first estimation activity, together with students' experience with value-bar graphs, forms the basis for visually estimating means. In this way, students can further develop qualitative and conceptual aspects of the mean such as intermediacy (somewhere in the middle), balancing and compensation (influence by all data values), and representativeness (it says something about the total, it accounts for all data values, and it can be used to compare two data sets).

4. Invent data according to aggregate features and coordination of center and spread in a meaningful context. Inventing data according to aggregate features of battery brands, or anything else, can stimulate students to develop an aggregate view of data. By answering questions such as what an unreliable brand with a long life span looks like, students learn to coordinate center and spread aspects. Students probably use predicates such as 'reliable' in different ways. Some may find brand D more reliable because its range is smaller, but others may find brand K more reliable because it has more of the same value. This reflects two different views on spread.

1. Estimate a large number in a picture. Implicit use of the mean. Discussion of 'average box', mid-range, median in other grids too.

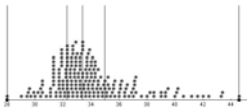


2. Compare different distribution aspects in value-bar graph. Informally discuss center, spread, and skewness aspects in suitable context (long life span, reliability, extreme values, position of the majority).



3. Explain what an average box is in a value-bar graph. towards a compensation strategy. Means represents and accounts for all values.

4. Invent data according to aggregate features. Coordination of center and spread in a context (e.g. unreliable brand with a long life span).

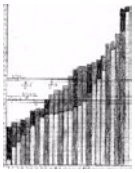


5. Compare distribution aspects of one data set in a value-bar graph and a dot plot. Necessity of concepts and structures underlying both representations. Discussion of spread: range and density views of spread. How are data spread out?

6. Estimate numbers with the mean and discuss sampling methods (e.g. balloon activity). Preparation to growing samples.

7. Characterize a larger sample.
Coordination of distribution aspects: mean more precise, more spread out, more around average, few low and high values, shape.

8. Represent various distribution aspects from a story.



9. Compare student graphs of one data set and reason about shape as an object. What happened to bump in the left graph? What happens with bump if sample grows? And if older students are measured?

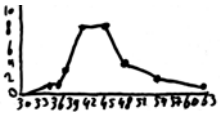


Figure 10.1: Schematic overview of the reconstructed HLT in grade 7

The first is a range view on spread and the second is a local view on spread that we characterize as a density view. In student language, this density view can be expressed as follows, in this example related to a dot plot organized into four equal groups: "Here the dots are spread out, but there they are close together." Students also combine the two views, for example when they explain how they examine a data set: "I first look at the highest and lowest, and then in between." It is probably necessary to introduce the notion of range to avoid situations in which students use the term 'spread' for the range only. Students are not likely to view spread as dispersion from a measure of center in such contexts as we have chosen, but this could be different in a context of repeated measurements of a 'true value' (see 10.4).

5. Compare spread of data sets in a value-bar graph and a dot plot. Describe how data are spread out with respect to organized dot plots. When describing how values or dots are spread out, students probably describe how the data values are distributed. This means that if the instructional design and teaching stimulate students to describe how data are spread out, they can also develop a sense of how data are distributed in relation to the meaning of the context. For instance, "the majority is more to the left" or "in the beginning the dots are less spread out." Students gradually develop a language in which they can express how data points are distributed, for example in reference to brand K: "in the beginning the steps are large and at the end they are small." The grouping options in Minitool 2 help students to express differences in density, especially in relation to four equal groups and fixed group sizes (for example in a problem situation such as the speed sign). After a few lessons with Minitool 2, students may be able to explain features such as the following: if the vertical lines of four equal groups or fixed group size are close together, the dots are bunched up or close together, even when data are hidden. This refers to a group of data with similar values. The implication is that they learn to see how the dots are distributed through the abstract diagrams that stem from grouping options in several ways and from hiding the data.

6. Estimate numbers using a notion of average, and discuss sampling methods. The balloon activity is a variant of the elephant estimation problem, but the sampling issue is more complex. In the elephant task, the population is the whole herd visible in the picture, but in the weight context, students need to use their contextual knowledge or simple sampling techniques to find out about students' and adults' weights. Again they may use average values as representative values and tools in their reasoning. The balloon activity forms the starting point of the growing samples activity in which students' initial guesses of weight averages are challenged. One disadvantage of the balloon activity is that it may contribute to the mean distractor effect.

7. Characterize a larger sample. After focusing on center and spread, the HLT aims at reasoning about shape as standing for the distribution of a data set as a whole. During the activity of growing samples, students predict the shapes of large samples (or even populations). In comparing their predictions with diagrams of real data sets, students are expected to reason about aggregate features of the samples and about shape. In Section 10.1.3 we elaborate on the growing samples activity and the importance of distinguishing between 'spiky' shapes of real data and idealized shapes with which experts model data.

8. Represent various distribution aspects from a story. The following example gives an impression of the extent to which seventh graders learned to reason about distribution. In the second assessment task of 1B, students had to draw diagrams that matched with a story on running trainings. From this item we concluded that most students were well able to symbolize case-oriented and many aggregate features of the story into the diagrams. Students can describe how data points are spread out or distributed, but reasoning about shape is probably difficult to accomplish in only twelve lessons.

9. Compare student graphs of one data set and reason about shape as an object. If students make various graphs of one data set they know well, they are likely to understand that these graphs are different representations of the same structure. Using such questions as mentioned in Chapter 8, teachers can support students in reasoning about shapes as objects. However, such reasoning is probably quite challenging to accomplish in grade 7 (see Section 10.1.2).

In short, we claim that students can learn to reason about distribution if an HLT similar to the one reconstructed above is used. This HLT offers an empirically grounded theory of how students may learn to reason about distribution. As mentioned before, an HLT always needs to be adjusted to local circumstances (Barab & Kirshner, 2001). It is likely, for instance, that more lessons are needed than we took, to compensate for the effect of the mini-interviews (see Section 10.3.1) and to avoid the pitfall of addressing too many topics without clearly defining what these topics are (10.5.1). Developing a well-defined terminology took more time than we had anticipated. Although the battery life span and speed sign activities had their merits in supporting particular types of reasoning, it is worthwhile to try out other contexts as well, because these appear to have disadvantages: in the battery context, students do not always expect variation and in the speed sign activity, students may focus on the speed limit as a cut point. Furthermore, the role of the median in the HLT has to be revised (see Sections 10.2.1 and 10.4).

10.1.2 Answer to the second research question

As motivated in Chapters 2 and 8, we used semiotics for analyzing the symbolizing process when students learned to reason about distribution. The semiotic theory most viable for this purpose turned out to be Peirce's semiotics. Compared to the theory of chains of signification, for instance, one major advantage of Peirce's theory of signs is non-linearity. Before summarizing an answer to the second research question, we briefly repeat the semiotic notions that are most important to the analysis.

In Peirce's theory, a sign stands in a triadic relation to an object and an interpretant. The interpretant is the response of an interpreter to the sign. Signs can have different functions depending on how they are interpreted. A sign is a symbol if its relation to its object and its interpretant is formed by habit or rule (and not by similarity for instance). A diagram is an icon representing relations. *Diagrammatic reasoning* involves three steps: constructing a diagram, experimenting with it, and reflecting upon the results. Anything that is thought or talked about is an object in Peirce's theory, and this object is mediated by a sign. From an educational point of view, it is therefore important that the topics of discussion are clear. The process of describing qualities of those topics or objects can be called *predication*. Next, *hypostatic abstraction* is one of the forms of abstraction that Peirce distinguished: a predicate becomes an object in itself that can have characteristics. This is linguistically reflected in the transition from a predicate (e.g. most, lying out) to a noun (majority, outlier). Symbolizing, within this theory, involves making a sign that is interpreted as a symbol, but generally also requires forming the object for which it stands: the symbol of a hill has to stand for an object, which students in general still have to develop (a notion of distribution). The notion of diagrammatic reasoning in combination with that of hypostatic abstraction offers a framework for analyzing the symbolizing process: symbolizing involves both the step of making a diagram (diagrammatization) and forming an abstract object such as distribution (e.g. by hypostatic abstraction). One advantage of using a differentiated notion of sign is that we can analyze students' difficulties with graphs in differentiated ways: we cannot simply say that histograms are difficult. Semiotically, interpreting elements of a sign such as reading off values from a plot is not so difficult, but to interpret the plot as a diagram standing for relations between data or even as a symbol standing for a frequency distribution requires much more conceptual understanding.

Earlier in this chapter we mention that the one-sentence summary of the HLT was to challenge students to reason about distribution aspects with increasingly sophisticated notions and diagrams. We can also conceive this semiotically as progressive diagrammatic reasoning about distribution aspects. In the remainder of this section, we highlight the most important ingredients of the symbolizing process when students learned to reason about distribution as an answer to the second research question.

Predication is a prerequisite for hypostatic abstraction. In all experiments, the estimation, battery, speed sign, and other activities were used to foster a process of predication: describing aggregate features of diagrams and what they represent. The activities aided students in describing features of the data sets with adjectives such as 'average', 'most', 'reliable', 'spread out', and 'low and high values' with respect to signs such as value-bar graphs and dot plots. The objects students talked about were mostly the bars and dots that stood for data values. The most common way of grouping data was into low, average, or high values (C1 in the Appendix). The instructional activities also offered opportunities for hypostatic abstraction, for instance of the average, majority, reliability, spread, and outliers. These notions have thus become distribution aspects with an object character, but these objects were still under development ('immediate objects', in Peirce's terms, not 'final objects').

Coordinating the steps of diagrammatic reasoning and diagrammatization according to aggregate predicates that are mean or spread related. One useful experimentation experience was using the value tool for estimating means in value-bar graphs (as could be done in Figure 10.2) and reflecting upon why this compensation strategy worked. Furthermore, student experimentation with the data sets in Minitool 1 and reflection on the features of the battery brands form the basis for diagrammatization of aggregate features such as "brand A has a long life span but is unreliable." In this way, students learn to think in center and spread-related terms with respect to diagrams, and to extend their case-oriented view with an aggregate view of data.

Diagrammatic reasoning about the bump. According to the HLT, the shape of a distribution had to become an object with aggregate properties. These properties are to represent features of the distributions, which in turn represent properties of the problem situation. To analyze this process and answer the second research question, we focused on the students' reasoning with bumps in class 1E.
The semiotic analysis in Chapter 8 shows that *diagrammatization* can involve multiple actions. Mike, for example, informally grouped the weight data values, used dots at certain positions to signify these groups and the frequencies, and connected the dots to one shape (Figure 10.1 right of item 9). Each of these actions has a history, either in the teaching experiment (e.g., grouping data and using dots) or in mathematics lessons (e.g., line graph).
During the *reflection* on the diagrams, the teacher used the term 'bump' to make this shape the topic of discussion. Students might have first interpreted the bump as an icon (a visual image of a bump), but the analysis shows that the meaning of the bump notion was not just iconic, and even changed from the eleventh to the thirteenth lesson, for several students at least. In the eleventh lesson, students used the bump notion to refer to a group of values close together in the middle part of the graphs. The

interesting thing is that the same data set looked so different in various student graphs. By relating the bump in Mike's graph to Emily's graph (Figure 10.1, left of item 9), the teacher stimulated students to specify what the object was which looked like a bump in Mike's graph and a straight line in Emily's. This object, a group of values that were close together, can be seen as a hypostatic abstraction. We can also speak of it as the common conceptual structure underlying aspects of both graphs. To conceive such a structure, however, students need to have a history of reasoning with such plots and similar phenomena (Bakker & Gravemeijer, 2003).

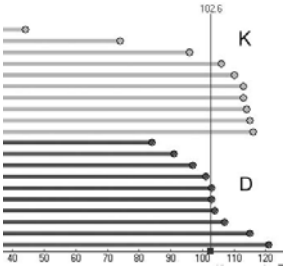


Figure 10.2: Reasoning with the 'bump' in a value-bar graph
("The bump of brand K is higher")

In the twelfth lesson, several students used the term 'bump' for a group of data even when there was no visual bump, for instance when they referred to the vertical straight part in value-bar graphs as a bump (Figure 10.2). This implies that the bump was not just a visual characteristic, but had become a conceptual object and even a reasoning tool, for instance for arguing which battery brand was better.

In the thirteenth lesson, several students referred to 'bump' as the whole shape, whereas before they had only referred to the area surrounding the peak of the hill shape, which for them represented a group of values being close together. The development of the bump as an object was probably stimulated by questions about hypothetical situations in which students needed the bump as an object. When we asked about the shape of a graph from a much larger sample, one student argued that it would grow wider if the sample were bigger because there would be more extreme values. Other students reasoned that the bump would stay the same because there would also be more 'average' values. This suggests that those students had an image of the distribution independent of the sample size. Students also answered the question of what would happen if students of a higher grade were measured. One student said that the bump would be shifted to the right. In that sense, the bump had become an object encapsulating the data set as a whole. Several students were able to relate aspects of that shape to distribution aspects such as average, majority, groups of values, and several acknowledged the stability of the shape across sample size. They even hypothesized on the shape of a large sample, which means they modeled hypo-

thetical situations with a notion of distribution, which was indeed the end goal of the HLT.
The analysis presented here can be taken as a paradigmatic example of explaining symbolizing as embedded in a process of diagrammatic reasoning in which opportunities for hypostatic abstraction occur. We anticipate that modeling in mathematics and science education can be analyzed as diagrammatic reasoning as well.

10.1.3 Answer to the integrated research question

This section provides an answer to the integrated research question in grade 8 in a more general way than Section 9.8. The eighth-grade teaching experiment was organized around the recurring activity of growing samples to test the conjecture that students could develop a notion of distribution by reasoning about growing samples. The integrated research question was how students with little statistical background can develop a notion of distribution by diagrammatic reasoning about growing samples. As before, we cast the reconstructed HLT in terms of anticipatory conjectures that can be tested and revised in practice. We focus on the lessons in which activities were carried out related to growing samples.

1. Diagrammatization according to group characteristics of larger samples. Because variability is the most fundamental concept in statistics, we argue it is important to choose contexts in which students are likely to acknowledge variability: without such understanding there is no need for taking a sample, computing a mean, measuring spread, or looking at the shape of a distribution. Industrial contexts such as the life span of batteries appear to be unsuitable starting points. The experience in grade 8 demonstrates that much can be revealed about students' intuitions of statistical notions when they design their own methods of sampling. Students presumably choose sample sizes that are too small or want to test the whole population. To challenge those views and to promote attention for aggregate features of data sets, teachers can ask what a larger sample with a specific aggregate feature might look like. This promotes diagrammatization according to group characteristics of data sets but also mental experimentation (a what-if attitude) and reflection on aggregate features and sample size. We asked, for instance, what a larger sample of a good and a bad battery brand would look like but, as mentioned earlier, we would not choose such an industrial context again.

2. Extend the samples to populations and create the need for drawing continuous shapes. In short cycles of growing samples of a familiar context, for instance weight or a less sensitive context, students may be asked to predict diagrams of samples with a specific size and compare those with real samples of that size. In this way, reflection can be stimulated about the diagrams and conceptual aspects of the sam-

ples in terms of center, spread, and shape. For larger samples students can use dot plots or continuous sketches and predict the shape of the population distribution. Hill, bell curve, pyramid, and semicircle are among the many possibilities. It depends on the context whether students will acknowledge the skewness of unimodal distributions. In the weight context, students probably do not expect a skewed shape, but skewness can be made a topic of discussion by discussing left and right limits in relation to the mean (as in Section 9.4).

When students reflect on comparing predicted diagrams and real data sets, it is crucial that distribution aspects become clear topics of discussion, so that objects can be formed that can be refined during the remainder of the instructional sequence. In particular, we think of the following distribution aspects: average, low and high values, outliers, range, spread, and shape. Linguistically, transitions should be stimulated from predicates such as 'most', 'lying out', and 'spread out' to nouns such as 'majority', 'outlier', and 'spread'. The formed hypostatic abstractions can be predicated again: for example, "the majority lies between 23 and 35, there is an outlier at 107, the spread is large." Drawing on their context knowledge, students can use these statistical objects as reasoning tools about shapes. However, care should be taken that students do not just mimic their teachers in using particular nouns.

The challenge is to strike the balance between providing space for exploration, participation, and reinvention on the one hand, and guidance towards culturally accepted and precisely defined notions on the other. It can be demanding for teachers to see the potential of students' ideas and supporting students in the next step. At some point, it will be necessary to discuss conventional notions such as range, median, mode, and outliers to avoid confusion between range and spread; between mean, midrange, and median; and between extreme values and outliers. This is preferably done after students have experienced the need for such distinctions themselves.

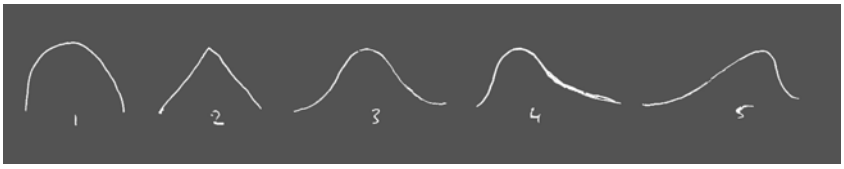


Figure 10.3: Shapes discussed in grade 8

3. Discussion of shapes. To address the distribution aspect of skewness, the different shapes students propose themselves probably need to be expanded with skewed shapes. In grade 8, students discussed the shapes of Figure 10.3 and used statistical notions such as mean and outlier to explain why certain shapes could not represent a weight distribution. During the diagrammatic reasoning about this, the statistical

notions of mean and ‘outlier’, were used as reasoning tools. Students also implicitly reasoned about frequency and density in this phase. To make skewness a topic of discussion as well, we introduced two skewed shapes. Students finally learned to describe shapes and distributions as being uniform, normal, skewed to the left or to the right. This means that shape had become an object that they could reason about with appropriate names.
The next step, which we overlooked in the eighth-grade teaching experiment, is to support students in recognizing shapes in dot plots or other plots in which variation around the smooth curve makes it difficult to perceive the signal in the noise. It is likely that students do not see the same shapes as experts, because they need to develop conceptual structures to perceive those or model the situation with a well-known distribution such as the normal distribution. They also need to experience when it is legitimate to reduce the complexity of a data set with a model (the signal) that includes variation (noise) around it. This implies that also idealized shapes should be made topics of discussion as well as deviations from such shapes, for instance by asking what different data sets (with the same distribution) have in common.
It seems wise to use both case-value plots (such as dot plots) and aggregate plots (such as continuous shapes). Consider this example. Most students in the final interviews could indicate where the mode and mean were in a sketch of a right-skewed distribution, but mistook the median for the midrange. By mentally going back to the two-equal-groups option in Minitool 2, several students could correct themselves and indicate where about the median would be in the continuous sketch. In this way, the measures of center were treated as characteristics of a distribution and not just as outcomes of computations performed on individual data values.

Having sketched the most important steps in the reconstructed HLT, we now discuss students’ notions of distribution. One of the end goals of the HLT was that distribution would become an object-like entity. In Chapter 5, we argue that a distribution is more like a composite unit than an object with a procedural and structural side and wonder what the procedural side of a distribution would be. In describing what a distribution was, students often used the term ‘spread out’: distribution is how the dots are spread out over the graph. From the way students talked about distribution, in particular during the final interviews, we inferred that they imagined the process of distributing dots over the variable as if growing a sample in a dot plot. We conjecture that such a process view of distribution could well be the procedural side of the concept, but we realize that our focus on growing samples has also fostered this view of distribution.
The analyses in Chapters 8 and 9 show that the reification process of distribution is in fact a complex process that involves many steps of hypostatic abstractions (cf.

Sfard, 2000a). Distribution is a multifaceted notion, the understanding of which requires understanding key aspects such as center, spread, density, and skewness. There even seems to be a reflexive relationship between the development of such characteristics of a distribution and the notion of distribution as an object or a shape: by reasoning about the occurrence of low, average, and high values, students expect a particular shape, and by reasoning about shape, students develop the meaning of distribution aspects such as mean, spread, density, and skewness.
We cannot answer the question of whether distribution had indeed become an object for the majority of the students without specifying what we mean by distribution and by object. The research has in fact yielded different levels of understanding distribution of which we mention a few.

1 We assume that students, before instruction, know what is typical and what is not about specific contexts, and that typical values occur more often than exceptional values. This forms the basis for students' reasoning about distribution in natural contexts. However, students mostly lack the language in which they can express such intuitions.
2 The activities in the beginning of the teaching experiments support students in describing various distribution aspects in informal and context-bound ways, such as the reliability of a battery brand in relation to diagrams (predication). Predicates such as 'most' and 'lying out' can become hypostatized as 'majority' and 'outlier', though the meaning of such terms is still under development.
3 Students then learn and use statistical terms for such aspects as range, spread, median, distribution shapes, and skewness in relation to various diagrams.
4 After five lessons, most students in grade 8 were able to express that there were few low and high values and many around average, which could be seen from the relative height in continuous shapes.
5 However, it appears to be much harder to recognize ideal shapes (as signals) within the noise of a 'real' data set as represented in a dot plot (lesson 8). Students reasoned *about* shapes (e.g. in lesson 6) but not *with* shapes as reasoning tools to solve other statistical problems. Moreover, they found it hard to predict the shape of data sets of hypothetical situations such as train delays.
6 Furthermore, there seems to be a difference in viewing a distribution as a shape that emerges from growing a sample, the presumed procedural side of the concept, and distribution as a statistical object with characteristics such as range, spread, mean, median, and mode. In the latter case, shape is an object that can be mentally manipulated and that can be used to predict and model new situations. Students conceiving a distribution in the former way need not understand it in the latter way.

Though students in the teaching experiments did not show understanding of all those levels, we assume that their diagrammatic reasoning experience forms a fruitful intuitive basis for the more technical applications of the normal distribution they will encounter in grades 10 to 12.

10.2 Other elements of an instruction theory

In the previous section, we presented answers to the research questions. These answers were related to the HLTs we used. In our view, HLTs are the most important ingredients of an instruction theory, but there are also more general issues belonging to a developing instruction theory. In the present section we discuss such issues for an instruction theory for early statistics education. We start with the key concepts except distribution because it has already been discussed extensively, discuss the most important diagrams, and finally use the notion of progressive diagrammatic reasoning to integrate students' development of key concepts, diagrams, and language.

10.2.1 Key concepts

This section provides a summary of the most important findings of the present research concerning the key concepts, independently from a specific HLT.

Data

Moore (1990) characterizes data as numbers in context. If there is no close connection between data and context, two things can go wrong. First, if students only see a data set as a batch of numbers, they might be inclined to conceive statistics as 'number crunching'. A possible consequence is that they calculate the mean whenever a question sounds statistical (in contrast see the motto of this chapter). The second thing that can go wrong is that students neglect the data and reason from the context only (as happened in the exploratory interviews). This implies that the norm should be established that the available data should be used when answering a statistical question. The teacher plays a crucial role in establishing such norms and practices. Students should come to understand why they need data and why these data should be created in a proper way to come to an appropriate conclusion. In line with the results of the Nashville team, we stress the importance of talking through the process of data creation as necessary preparation to seeing data as numbers in context. In fact, talking through this process is also a way to address the measurement and sampling issues: what variable exactly is measured and how? However, such guided discussions alone may not suffice; in our view, students should also experience a whole investigative cycle from asking a question, collecting their own data,[58] analyzing

58. A survey in grade 8 showed that students favored the activity of collecting data on car colors because they liked the context and liked doing something.

data to communicating the results and perhaps refining the question and the data collection.

Center

Students can learn to calculate means and medians without too much effort, but this does not imply that they perceive means or medians as measures of center. There is considerable research showing that students generally do not see those values as group descriptors (for an overview see Konold & Higgins, 2003). We have offered remedies for learning the mean: in Chapter 6 we show how students can be supported in coordinating their computational knowledge of the mean with their intuitions of average. The estimation and compensation strategy proved useful for making this connection.

However, the median turned out to be conceptually even harder than we had expected. Like Cobb, McClain, and Gravemeijer (2003), we observed that the median can have two distinct meanings. The first, finding the median of a set of data points, is relatively easy; but the second, the median as a representative value or a characteristic of a distribution, is difficult to develop. One of the difficulties we identified was the following. For the students in the present study, the mean accounted for all the data points, but the median did not. One of the reasons is probably that the median is independent from the values of the differences. Therefore, many students viewed the mean as more precise. It seems especially counterintuitive to students to take a median, which is just order-dependent, in a representation with a rational scale (such as a dot plot).

In search of alternatives for learning a notion of center and measures of center, we have come to the hypothesis that student intuition of an average group or a modal clump can be used to develop a notion of center. Konold and colleagues (2002) argue that such 'modal clumps' can function as measures of center (and spread), and that these mostly encompass mean, median, and mode (at least in unimodal distributions). We assume that it is fruitful to let students indicate the ranges of where they see clumps in the data (Figure 10.4) as a precursor to viewing that part of the data as the center and later using formal measures to locate its position. In particular, the median could be a useful measure of where the clump in a skewed distribution is.

Another way to support students in developing a notion of center, which is not in the spirit of exploratory data analysis, is by focusing on true values and errors in repeated measurements. This is how the mean and normal distribution have been developed historically, and a few researchers have taken this path. Petrosino, Lehrer, and Schauble (2003), for instance, let students measure the height of a flag pole, the length of a pencil, and the height of model rockets in their flight. The activities helped in fostering a notion of a true value (median and mode) and variation of errors. In line with this idea, Konold & Pollatsek (2002) reason that the ideas of signal

and noise are useful conceptual underpinnings for seeing an average value as a measure of center or a true value. The true value approach seems to ask for scientific contexts and the clump approach for contexts in which there is no true value. Both contexts appear to have their advantages and disadvantages, but care should be taken in combining them: as is clear from the history of statistics (e.g. Porter, 1986), the transition from variation in a measurement error context to natural variation (such as with height) was quite a leap conceptually.

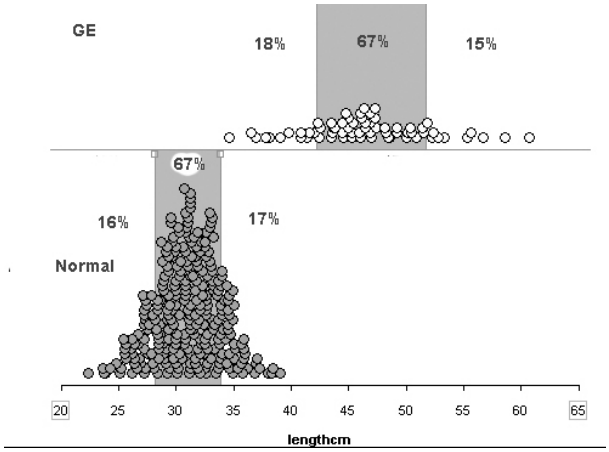


Figure 10.4: Clumps indicated with dividers and percentages (in Tinkerplots); lengths of normal fish and genetically engineered fish (hypothetical data set). The 'clump' of the genetically engineered fish is more to the right (higher values) but also has a wider range.

Spread

In recent years, researchers have come to call for more attention to the notions of spread and variation across all grade levels. The stress on the formal measures of center has probably kept spread in the background and, as a consequence, little is known about students' notions of spread and variation (Meletiou, 2002). One of the challenges of teaching spread is that the only easy way of measuring spread is by the range of a data set (historically the first measure of variation). This, however, is a rather crude measure of spread, just as the midrange is a crude measure of center. The next candidate is the interquartile range, but the present research shows that quartiles take much time to develop (except perhaps for a few *vwo*-students). One of the less formal ways to reason about spread is with the four-equal-groups option in Minitool 2 as a precursor to quartiles and box plots. If a computer tool makes such divisions, students need not be bothered with the precise definition of quartiles. A disadvantage is that students, along with their teachers, may think that exactly 25% of the data values will be in each part, whereas this percentage is rarely exactly 25%. It turned out that the transition to viewing spread as dispersed from a mean or a me-

dian was a big leap for these students. For them there seemed to be no reason in the contexts we used to measure differences from the mean, as is necessary for standard deviation. In the context of repeated measurements, this way of conceiving spread probably makes more sense to students, because they can think of possible causes of errors. For their fourth graders, Petrosino, Lehrer, and Schauble (2003) used a so-called 'spread number', which was defined as the median of the absolute differences with the median of the data set.[59]

Sampling

As in the case of spread, sampling receives little attention in most middle school curricula. This may be caused by the lack of numerical calculations that can be trained and tested. Moreover, statistics is taught as part of the mathematics curriculum, which generally focuses on well-defined notions and calculations rather than on statistical reasoning. Another reason could be that sampling is a difficult notion for students to develop. Yet it is important to address the sampling issue, for instance to make sure that students understand how the data were created. The activity of growing a sample is an intuitively reasonable way of addressing sampling. Resampling can later be used to address the variation between samples of the same size.

The present research confirms various results of other studies (e.g. Watson & Moritz, 2000). For example, students in our study indeed did not specify the selection of items and too easily assumed that taking 'a few' gives a fair image of what the question was about. Yet, a little instruction can change this: as it turned out in grade 7 and 8, it was relatively easy to promote the insight that larger samples generally provide a better image but are also more time-consuming and costly than small samples. On a concrete level, students were sensitive to bias: they understood that counting car colors before a (Dutch) post office would yield too high a percentage of red cars. However, promoting the insight that there is variation between samples of the same size proved difficult in grade 8.

10.2.2 Diagrams

In this section we discuss the results concerning various diagram types: value-bar graph, dot plot, histogram, and box plot.

Value-bar graph

Value-bar graphs are not commonly used in statistics and they are only useful for relatively small data sets. Yet there are several reasons to use value-bar graphs in a middle school statistics curriculum. First, students easily interpret value-bar graphs,

59. Assume that 9.1, 9.3, 9.7, 9.8 and 10.6 are the measurements of the height of a flag pole. The median is 9.7 meter, and the absolute differences are 0.6, 0.4, 0.1, and 0.9. The median of these four absolute differences is 0.5 meter, the spread number.

probably because they are already acquainted with bar graphs (with categorical data). Second, the present research shows that value-bar graphs can help students estimate the mean visually and understand the connection between the computation of the mean and its qualitative aspects such as intermediacy, balance, and representativeness. We assume that value-bar graphs provide a better model than the balance model because students at this age probably do not know the physical laws of balance (cf. Hardiman et al., 1984; Pine & Messer, 2000) even though they might have some experience with see-saws. Third, letting students compare different representations of distributions turned out useful; one of the successful comparisons was that of value-bar graphs and mound-shaped graphs (Chapters 6 and 8). Fourth, if students have not already developed a notion of variable, experimenting with the value tool in Minitool 1 and reasoning about the endpoints of value bars may help them develop such a sense of a variable (Gravemeijer, 2000b). Fifth, when students are asked to make a graph of a data set, many of them make value-bar graphs, mostly with vertical bars. We conclude that value-bar graphs are useful in a statistics unit at the middle school level.

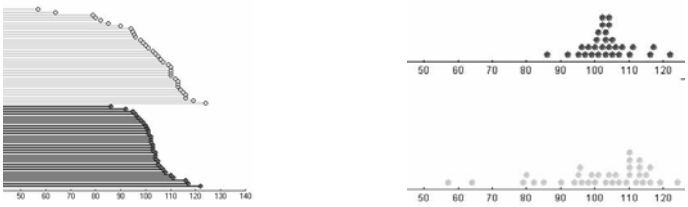


Figure 10.5: Value-bar graph in Minitool 1 and a dot plot in Minitool 2, both of the battery data set used in classes 1B and 1C

Dot plot

Like value-bar graphs, dot plots are easy to read for students but not often seen in statistics textbooks. An advantage of dot plots is that they can represent much larger data sets than value-bar graphs. Though dot plots are sometimes used by professional statisticians (Scheaffer, 2000; Wilkinson, 1999) and provided by commercial educational statistics software packages (Fathom, Tinkerplots), we also heard criticism on dot plots: a dot plot has no vertical axis, which means that the height of a dot has no formal meaning, unless dots are stacked as in Figure 10.6. In an introductory course, in which there is no way of discriminating between frequency distributions and probability density functions, this may also surface as an advantage: students have to construct a meaning to the height of the graph and they can come to see it as an informal measure of density. As a consequence, the shape of a distribution is oftentimes more visible from a dot plot than from a histogram.

We conclude that the dot plot is a useful graphical representation in an introductory statistics course as it allows students to see the individual data points.

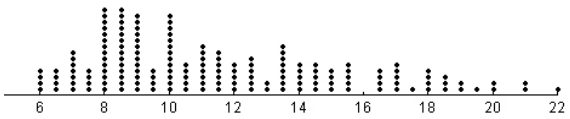


Figure 10.6: Dot plot in which dots are neatly stacked due to rounded values and small dots; thus, height represents frequency

Histogram

The histogram is one of the most commonly used diagrams in statistics. However, it is common knowledge that many students mistake histograms for bar graphs (Baker et al., 2002). For instance, many students misinterpret the height of a bar as the height of people instead of a frequency. This mistake also occurred in the present research in grade 7, despite preparation with dot plots and the fixed-interval-width option in Minitool 2.

In the rationale of the Minitools, there was a learning trajectory from value-bar graphs, to dot plots in which students can organize data points with their own groups and with fixed intervals. Once fixed intervals are chosen, students using the revised version of Minitool 2 can select the histogram option and hide the data points. The jeans activity (7.12) was intended to motivate the choice for fixed intervals, but in grade 7 it did not lend itself so well. In both grades, students preferred very different plots for solving such problems as the jeans problem. In grade 8, some students preferred making their own groups even after we had shown them the options of fixed intervals. We concluded that for them the interval option was not transparent, let alone the histogram option. There were students who preferred to keep the dots in their plots whereas others found the histogram and box plots options without the dots "less busy" or "better organized". This leads us to the conclusion that providing the full range of possibilities gives all students the chance to use a representation they can use as a meaningful reasoning tool (cf. Treffers, Streefland, & De Moor, 1994).

We advise against introducing histograms in early middle school grades for the following reasons. Students in grade 7 needed much time to develop a sense of center and spread, and this took away time allotted for learning the ins and outs of histograms and box plots. Of course, it is not difficult for students to read off particular values from histograms or box plots, but reading off values (elements in a complex sign) is semiotically seen as quite different from interpreting a sign as a diagram (representing relations between data values) or as a symbol standing for one object (for instance of a normal distribution or a frequency distribution). As we argue in Chapter 2, we do not see much use in teaching students notions (mean, median) and graphs (histogram, box plot) that they can only interpret on a superficial level (albeit in an exact way) and not use as meaningful reasoning tools for analyzing data.

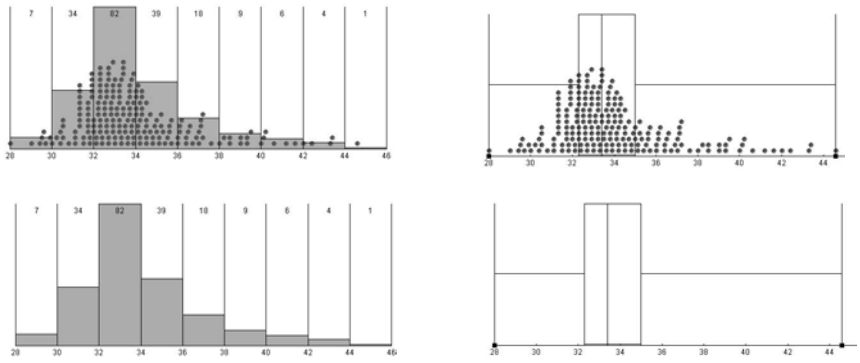


Figure 10.7: Histogram and box plot options in Minitool 2 (without and with hiding data)

Box plot

Box plots are useful representations for summarizing and comparing distributions. For students, however, this type of plot incorporates several statistical notions such as median and quartiles, which are conceptually demanding (Friel et al., 2001). It is therefore no surprise that students have troubles with box plots, even at the high school level (Konold et al., 1997). The pitfall of introducing box plots at the middle school level is that precious time is spent on how to make box plots and interpret superficial features, for instance what the median is. In our view, this time is better spent on doing statistical data analysis with simpler tools that are more likely to become meaningful reasoning tools. We therefore propose not to use box plots at the middle school level. For the high school level, routes as proposed in Chapter 2 can be used (Figure 10.7).

10.2.3 Progressive diagrammatic reasoning

The notion of diagrammatic reasoning turned out to be useful in capturing the integrated development of key concepts, diagrams, and student language. We have added the adjective 'progressive' to the notion of diagrammatic reasoning in analogy to 'progressive schematizing', which is commonly used in the RME theory (Gravemeijer, 1994; Treffers, 1987). What can we learn from the results for an instruction theory for statistics education? In short, it is pivotal to stimulate the steps of diagrammatic reasoning: diagrammatization, experimentation, and reflection. In our view, these are three core steps in learning statistical reasoning. These steps need not occur in a particular order and, in stressing the importance of those three steps, we do not intend to be exhaustive (cf. Wild & Pfannkuch, 1999). In the remainder of this section we discuss these steps in more detail.

Diagrammatization

For the step of making diagrams, we distinguish cases in which students make their own diagrams and cases in which students are offered diagrams, for instance in computer software.

Throughout the research we have noted that it proves productive to let students make their *own diagrams* (cf. the third RME tenet of stimulating students' own constructions). Students' own diagrams are likely to be meaningful to them and the variety of diagrams can be used to discuss different aspects of the data sets or diagrams (see also DiSessa, 2002). However, apart from offering opportunities for reinvention, there also needs to be careful guidance. Otherwise, students might draw many bar graphs, but not many other plots. If students make their own diagrams, they are likely to use diagrams they know, because habits are formed during previous experimenting with them (the interpretant is a response to a sign, and this response is often habitual[60]). Mike's diagrammatization of his bump, for instance, involved grouping data, symbolizing them with dots, and drawing a line through those dots, which are all habits formed in previous experiences. Emily, for instance, symbolized the mean with the value tool with which students in previous lessons had estimated means.

If diagrams are *offered*, it is important that these diagrams are well prepared and relatively easy to interpret. Otherwise, they are unlikely to become tools in students' reasoning and problem solving. The value-bar graphs and dot plots offered in the two Minitools proved relatively easy for students to interpret as representations of data values and relations between those values (such as "few low and high values" or "many around average"). Much of the research literature on visualization and representation turns out relevant to this aspect of diagrammatic reasoning (e.g. Goldin & Janvier, 1998; Janvier, 1987; Kadunz, 1998; Presmeg, 1998).

Experimentation

Experimentation with diagrams can be done physically and mentally. Computer software can be useful for *physically* experimenting with diagrams. Students get the opportunity to dynamically interact with a data set and make different diagrams of that data set. The advantage of this possibility of exploring data sets over learning how to make specific plots in a procedural way is that students can experience that one data set can be represented in different ways and that different features of the data set can be deduced from different diagrams. However, such experience does not emerge automatically. For instance, though histogram and box plot options were prepared in grade 8 with computer tool options and instructional activities, only few students found those representations helpful. More lessons were required for discussing the merits of the different representations than were available.

Apart from physically experimenting with diagrams, students can also *mentally* ex-

60. Cf. the development of instrumentation schemes (Drijvers, 2003, p. 98).

periment with them. This ability is important for being able to choose a suitable representation of the data and anticipate what data look like in a particular type of diagram. In the present research we tried to promote mental experimentation in different ways. One was by letting students report their findings in the role of data analyst (as opposed to just solving a statistical problem). This probably helps them imagine what a data set would look like in a specific type of diagram. Another way of promoting mental experimentation was by asking questions that could promote a 'what-if' attitude. We asked, for instance, what a graph would look like if the sample were much bigger and what would happen to the bump if an older group of students were weighed.

Reflection

The best reflection that we have observed was during whole-class discussions and mini-interviews. It proved particularly difficult to organize reflective discussion in a computer lab.[61] We highlight two important ingredients of the reflection phase: stimulating predication and creating opportunities for hypostatic abstraction.

Predication is the process of describing features of diagrams or what they stand for. By using suitable data sets of appropriate problem situations, students can be guided in the type of predicates they use. In the present research, students predicated values or groups of values in terms of 'lying out', 'staying behind', 'most values', 'the average box', 'high values', and so on. Furthermore, they predicated dots as 'close together', 'bunched up', 'packed', or 'spread out'. They also talked about data sets as consisting of low, average, and high values. The importance of predication is that specific aspects of diagrams can become topics of common attention ('objects') and that distinctions can be made, for instance between range and spread.

One of the ways to create opportunities for *hypostatic abstraction* is to ask students what they precisely mean by terms, for instance during class discussions. It is advisable to use questions that we asked in the mini-interviews, such as the following. What do you mean by spread? What exactly do you mean by 'average'? What do you call 'low values'? What would you take as the 'average box' in this case? Where do you see the majority? In this way, students develop a vocabulary with which they can more clearly define the objects they talk about. Sometimes it is possible to introduce an official name for what students already describe: the median for the middlemost value, the midrange for the mean of the minimum and maximum values, and the mode for the value that occurs most often. In the present research there have been several instances of such reinventions. More common, however, is the necessity to deliberately introduce notions such as range to avoid confusion.

Another way of stimulating hypostatic abstraction is to create situations in which predicates can be used as objects. When the teacher in grade 7 asked what would

61. It would help if the teacher could switch off the screens with one button.

happen with the bump if older students were weighed, several students could operate with the bump as an object by shifting it to the right. Furthermore, it is useful to compare two diagrams of the same data set (such as Mike and Emily's graphs). We can explain this in the following way. If the diagrams, or aspects of them, look different whereas students know that they represent the same data set, there could be a cognitive conflict that needs to be explained (e.g. bump versus a straight part in the diagrams). Then there may be a need, from a student perspective, for a conceptual object that underlies the two diagrams (majority, bump).
This step of reflection links well with research on communication (Cobb et al., 2000; Steinbring, 2000), discourse (Sfard, 2000b), interaction (Dekker & Elshout-Mohr, 1998; Nelissen, 1987), and the proactive role of the teacher in class discussions (McClain, 2000).

10.3 Discussion

In this section we discuss the following topics: methodology, RME design heuristics, symbolizing, computer tools, and a comparison with the Nashville results.

10.3.1 Methodology

The methodology chosen for the present research as to contribute to an instruction theory for statistics education was design research. In this section we discuss a few of its merits and constraints.
By staying close to educational practice (the study was carried out in regular classrooms with their regular teachers), the research gains ecological validity. This implies, as Cobb, Confrey, et al. (2003) write, that the results are already filtered for instrumental effect. Moreover, design research may well help in bridging the gap between theory and practice. However, we sometimes felt the need to gain more certainty about particular issues. For instance, we became interested in the question of whether it makes a difference to students if value bars are presented vertically or horizontally when estimating means or finding medians. Comparative empirical research, within the setting created by the design research, could yield more insight into such questions, but this requires a larger team of researchers. Likewise, we could not reach the same rigor in analyzing students' notions of spread as for instance Watson and Moritz (2000) do in analyzing students' notions of sampling using clinical interviews. Such large-scale interview studies are rich sources for what can be expected when we start instruction to students of a particular age. Nevertheless, they do not prescribe how students' notions can be improved. For instance, we also encountered 'small samplers' in our study, but it turned out relatively easy to convey that larger samples are generally preferable to small samples, but that drawing larger samples takes more time and money. One merit of design research is that it can be used to analyze the development of students' notions when we try to support that development (cf. Frick, 1998). We propose combining design research with

small comparative studies to gain more certainty on crucial decisions within the design (cf. Brown, 1992).

Another issue that reveals both a merit and a constraint of design research is the generality of its results. We contrast it with a method common in educational research to arrive at general results: pose general but dichotomous questions and work those out in concrete situations (e.g. Milo, 2003). One such dichotomous question is whether we should provide students with a model or let them design one themselves (for a discussion see Van Dijk et al., 2003). A significantly better score in one condition may lead to a general result. In the present design research we worked the other way round. We developed activities through a cyclic process of design, trial, analysis, and revision and arrived at activities that worked out well in a specific content domain. We then tried to generalize the results by describing patterns in student learning and the means of supporting that learning. In this way, a subtle answer is given to the general question of whether to provide students with models or let them design diagrams themselves. We should do both, and the interplay is a subtle balance (e.g. in the growing samples activity in Section 9.4). This is where we need domain-specific instruction theories, also with respect to computer tools. As Hoyles and Noss (2003) observe: small changes in technological tools can have major influences on students' learning. Dichotomously set up comparative research does not appear to offer the small grain size to explain the effects of those seemingly small changes in design.

Last we discuss how generalizable or replicable the results of the present research are if we take a learning effect of the mini-interviews into account. The mini-interviews stimulated reflection. Due to the questions we (including the assistants) asked, students were invited to predicate features of diagrams and to explain what they meant by particular words. It is likely that the mini-interviews have not only accelerated the learning process but have also improved its quality. The mini-interview questions were formulated before the lessons and can therefore be taken as part of the HLT. If the HLT is to be used again, the designer and teacher should make sure that questions asked in the mini-interviews are also asked in the instructional materials and during class or small-group discussions.

10.3.2 Heuristics of Realistic Mathematics Education

In this section we reflect on the RME heuristics of historical and didactical phenomenology, and guided reinvention (2.1). What has the historical phenomenology contributed to the design? Many of the historical hypotheses functioned as elements of the evolving HLT. Of the 28 hypotheses we formulated in Chapter 4, we found empirical support for 11. One, H10, needed refinement: the Trial of the Pyx turned out unsuitable because the historical context became an extra obstacle to overcome (Van Amerom, 2002, makes a similar observation). We have not been able to test the remaining 14 hypotheses, which were mainly about median and distribution.

More specifically, the most apparent success was the idea to start the HLT with estimation to support reasoning about average and sampling (H1, H6). The elephant and balloon activities were directly inspired by the historical study. Next, we would probably not have been sensitive to students' use of the midrange as a possible precursor to the mean if we had not known that the midrange was one of the possible precursors of the mean. Instead of avoiding the midrange, as happens in all textbooks we know, we took advantage of its turning up in discussions and used skewed distributions to challenge its use. Thus students became more sensitive to how data values were distributed.

More generally, the historical study turned out to be helpful for looking through the eyes of the students. Instead of taking the precisely defined historical end products as our framework, our own statistical concepts became more 'fluid' again. As designers, we sometimes need to go through a process opposite to reification. This 'deconstruction' process is a first essential step of a didactical phenomenology.[62] The next step is to reconstruct a revised and improved version of the historical learning process that young learners might go through while using the knowledge they have and their ancestors did not have (Section 4.1).

Using the lignification metaphor by Frege, the motto of Chapter 9, we re-address the issue of guided reinvention. On the one hand, we need to provide students enough opportunities to think freely and creatively. This implies that it is beneficial to let students invent their own diagrams and ideas. Activities with a playful element such as the growing samples activity can succeed in engaging students in statistical reasoning (Sections 9.3 and 9.4). Students can work at their own level: in making a diagram they can think of the statistical notions and meanings as they have at hand (cf. final test in 1B). For the software that is used, this means that it should be easy enough for students to make sense of what the software can do. This all comprises the "soft and sappy" reinvention part of guided reinvention.

On the other hand, students need to be guided carefully: their reasoning needs to be formalized, and their notions have to become more abstract and general. Apart from the instructional activities, the instructional software has to be designed carefully. They must provide enough structure and direction to afford level raising (emergent modeling, abstraction, generalization, formalization). And probably the most important guidance comes from the teacher. This is the 'guided' part of guided reinvention. Our experience is: the more autonomy we want to allow students, the more we have to invest in planning (cf. Bakker & Gravemeijer, 2003).

62. By 'deconstruction' we do not mean task analysis, which aims at decomposing tasks into smaller and easier steps, but 'liquefying' the historically reified concepts.

10.3.3 Symbolizing

In this section, we discuss the advantages of using semiotics for analyzing students' symbolizing process, and in particular of using Peirce's semiotics as opposed to chains of significations. Some readers may ask what using semiotics as an instrument of analysis adds to using common sense. By framing episodes of student learning as examples of more general issues such as diagrammatic reasoning or hypostatic abstraction, the results can be generalized and the insights can be applied in new situations. For example, insights into students' diagrammatic reasoning can be linked with insights into students' ways of modeling in mathematics and science education. In this way, the present study gains theoretical validity and contributes to theory development. Apart from this issue of generality, our experience is that the semiotic framework used provided insights that we had probably not gained without it and provided a vocabulary to express insights more precisely than without those notions.

In Chapter 8 we argue why Peirce's triadic sign notion better suits the purpose of analyzing such non-linear processes as learning statistical reasoning than the dyadic sign of Saussure or Lacan. What makes Peirce's notion of sign non-linear is that the interpretant can be the reaction to multiple signs and that it can be the production of multiple new signs. Several researchers have used the notion of chain of signification for analyzing processes of symbolizing. In Chapter 8, we conclude that this notion is too limited for analyzing more complex learning processes, for example, if they compare signs. Instead, we use the notions of diagrammatic reasoning, predication, and hypostatic abstraction to analyze student learning. The adjective 'progressive' was added to diagrammatic reasoning to stress that it should lead to, in this case, reasoning about distribution aspects with increasingly sophisticated notions and signs. Note that it is possible to describe chain-like signification processes with the Peircean notion, because the interpretant can be a new, more complex sign, the interpretant of which can again be a new, more sophisticated sign, and so on. We therefore regard Peirce's theory as superseding (in the sense of Steffe & Thompson, 2000) the theory of chains of signification.

Another advantage is that the notion of symbolizing can be embedded in a superseding framework of diagrammatic reasoning. Symbolizing is literally making a symbol, and this includes both making the sign, for instance by diagrammatization, and forming the abstract object it represents (e.g. by hypostatic abstraction).

One feature of Peirce's semiotics that could be seen as a limitation is its lack of grounding in a psychological theory. Hence, researchers have combined semiotic and psychological perspectives. For instance, Seeger (2001, 2002) has compared the theories of Peirce and Vygotsky, and Van Oers (2000) has proposed a psychosemiotic theory which draws on both semiotic and psychological theories. Moreover, several authors of Anderson, Sáenz-Ludlow, Zellweger, and Cifarelli (2003) have contributed to a social semiotics as their framework. One further research challenge

we see is to develop a semiotics for IT applications. Semiotics has been developed for static signs, but signs in IT tools can be dynamic and interactive: if an interpreter dynamically interacts with an IT sign, how should the interpretant be defined? The interpreter's response is tied to the variety of rules of the IT tool that learners need not be aware of. One way to investigate this issue is to unravel the relationship between semiotics and the instrumental approach, which theorizes on how artefacts such as IT tools become instruments for solving mathematical problems in student hands (Artigue, 2002; Drijvers, 2003).

10.3.4 Computer tools

What can we learn from the experiences with the educational statistics software? In this section, we tentatively discuss a few affordances and constraints of educational software such as the Minitools and formulate hypotheses about different types of educational statistics software that have been developed for middle school students.

One attractive feature of the Minitools is that it hardly takes any time to learn the technical aspects of the tools, unlike with computer algebra systems (Drijvers, 2003). However, there are few opportunities for exploring different representations, which we view as a constraint. Throughout the research we have looked for ways to let students make their own graphs and compare different representations of the same data set. When students made their own visual representations, these were generally very similar to the Minitools representations. This implies that software can heavily influence the way in which students make diagrams of situations. It is striking that many students made vertical bar graphs though Minitool 1 only offers horizontal bars. This can be interpreted in two ways: the first is that the applet in its present form is too restrictive; the second is that students were not bound to the representation they had used. We assume the students in our study could have benefitted from using a more expressive tool (in the sense of Doerr & Zangor, 2000).

How can computer tools support diagrammatic reasoning? Computer software such as the Minitools appears most useful for experimentation with diagrams. For example, Minitool 1 with its value tool supported a visual compensation strategy in the present study (but not in the Nashville research). The experimentation experience with particular types of diagrams forms the basis for reflection. In supporting diagrammatic reasoning, computer tools should in our view also offer user-friendly options for diagrammatization.

The three Minitools form a series of small applications (or applets) that are designed for a particular HLT. We characterize such series as route-type software (Bakker, 2002). There are also larger applications such as Fathom, Tabletop, Tinkerplots, and VU-Stat that are not tied to specific HLTs, but that offer a landscape of possibilities for analyzing data. We call this landscape-type software in analogy to Fosnot and Dolk's (2001) notion of the landscape of learning. On the basis of our experience with the Minitools and Tinkerplots we have formulated the following hypotheses,

which can be generalized to route-type and landscape-type software.

1 Small applications such as the Minitools are useful for teaching specific issues, such as visually estimating means or preparing the introductions to histograms or box plots if there is no or little time to learn the software. However, there is a risk of a narrow path offering little exploration space for genuine data analysis.
2 It is easier for teachers to guide students in using simple tools and to discuss their reasoning with the tools because of the limited variety. Conversely, if students use a larger application in which they can make many different plots such as in Tinkerplots, it is more demanding for teachers to guide their students.
3 When using larger applications, students will spend more time finding a good representation of the data and on doing genuine data analysis, but perhaps learn less about specific topics that smaller applications can draw the attention to.
4 When using larger applications such as Tinkerplots, students need much time to learn specific features of the software and will not understand many of the plots they produce. Much reflection time is likely to be spent on the meaning of unconventional plots. Hence, students' meta-representational skills (DiSessa, 2002) might improve, but their knowledge about conventional plots could deteriorate when using large applications.

It might be sensible to start with simple tools if students are young and inexperienced in analyzing data, and use larger applications for students who already have some statistical understanding and skills in reasoning with various plots. We assume that the students in our study could well have coped with a larger application such as Tinkerplots. Moreover, Tinkerplots offers the option to gray out options so that smaller environments such as value-bar graphs or dot plots with limited grouping options can be presented to students.
The question arises of what criteria there are for selecting a computer tool. We suggest two. First, is the tool likely to become meaningful to students? Second, can teachers guide students in learning to reason with this tool? Whichever tool is chosen, the instructional activities, assessment, and the way of teaching should be in tune with the tool, and vice versa (Chapter 9; cf. Kanselaar et al., 1999).

10.3.5 Comparison with the Nashville research

In this section, we compare the present research with the Nashville research. After mentioning the practical differences we compare the most important results.
The Nashville research was carried out with one group of students over a two-year period (37 lessons in grade 7 with 29 students, and 41 lessons in grade 8 with 11 students). Because we could not get more than twelve lessons, we carried out several teaching experiments: four in grade 7 and one in grade 8 (between ten and fifteen lessons per experiment). The level of the Dutch *havo-vwo* students was probably a

little higher than that of the Nashville students: about 35-45% of the Dutch students attend the *havo-vwo* tracks of education, whereas the Nashville students joined a 'magnet' school which was attended by the top 50% (Cortina and McClain, personal communication, February 2, 2004). The Nashville research was carried out by a large group (Section 2.3), whereas the present research was not. The two teachers in the Nashville research were team members who only gave one lesson per day, whereas the teachers in our study also taught several other classes per day as part of their regular job.

In retrospect, we characterize our teaching experiments as less linear in different ways. We often asked students to make and compare their own graphs and to compare Minitool 1 and 2 representations. As a consequence, we needed another semiotic framework that would allow for network-like analyses of students' learning, especially when comparing representations. In Chapter 8, we argue why the Peircean semiotics best suited our purpose. The Nashville team had much more time to establish the norms and practices they wanted to establish in their class. Thus they were in a better position to be confident that the meanings of words were 'taken-as-shared' before moving to a more advanced issue or tool. This may explain why the Nashville team could describe the collective learning as a chain of signification. Because our situation more captured regular school practice and we had less time per teaching experiment, there was more variety between students within one class. The time span was too short to focus on group processes such as emerging norms and practices. Instead we were interested in students' conceptual development of center, spread, sampling, and distribution.

We now compare the main points of departures and results, which are indicated as P# and R# as in Chapter 2.

P2. We did not restrict our contexts to ones that were *socially important* such as AIDS and CO_2 emission, because that would cut down the number of possibilities drastically. Activities that students participated well in, such as the elephant estimation, the car color activity, and growing samples activity, were often not situated in socially important contexts.

P6. We did not only ask students to *compare distributions*, but also to describe and structure single distributions. One reason was to avoid vertical comparisons of slices of distributions. Another reason was we considered it dull for students to compare two distributions in each activity.

P7, R15. As we argue in Chapters 4 and 5, we did not want to avoid the *mean* and Chapters 6 and 7 show it was not necessary to do so. Moreover, the mean is probably the most used statistic.

P11 and R8. We have not stressed *multiplicative reasoning* because the Dutch students were already reasonably fluent with percentages. Moreover, we tried to avoid a situation in which they would vertically compare slices of two distributions or

would just compare percentages left and right of one value. For instance, we changed the question of the speed trap activity to avoid students from focusing on the speed limit as a cutting point.

R1. Talking through the *data creation process* indeed turned out very important to bring the context to life and address the sampling issue. Without a sense of the sampling issue, students often do not understand what the data stand for.

R2, R17. We have therefore paid more and more attention to *sampling.* Though sampling is a complex notion, it also has aspects that can be developed rather easily. For instance, students quickly came to understand that larger samples are mostly more reliable and several sampling and distribution aspects could be coherently addressed by discussing growing samples (Chapter 9).

R5. As in the Nashville research, the initial *case-oriented* views (reasoning about features of data points) were extended with *aggregate* views of data (reasoning about distribution aspects).

R6. Both in the Nashville research and the present research, students came to reason with *shape* as bumps and hills as symbols, not just visual images.

R9. In the present research, students learned to estimate means visually with the value tool, but in the Nashville research this was not the case. We see two possible explanations. First, the Dutch students already had a better understanding of the mean, and second, the activities of the first lessons in grade 7 enabled them to reinvent the compensation strategy.

R12. We found empirical support for the importance of the role as *data analyst.* For instance, when the battery problem was cast into the context of the *Consumentenbond*, the Dutch equivalent of *Consumer Reports*, students gave more and better arguments than in a factory context in which they were inclined to sell an advertising pitch.

R16. As with the Nashville team, we encountered the difficulty of designing good activities to support students' development of the *median as a measure of center*. In line with the Nashville team we assume that students should first have a sense of the center of a distribution before they can measure that center (e.g. a clump in the center of a data set) with a median, and use medians to compare distributions.

In short, the present study supports many of the results found in the Nashville research, but there are also a few differences. It is noteworthy that very similar patterns in students' learning occurred in activities such as the battery life span problems, although the two populations differed in level and educational context. We furthermore designed a few new activities that support students' developments of aggregate views, for example the data invention tasks, comparison of representations, and growing samples activities.

10.4 Towards a new statistics curriculum

In this section, we make suggestions for a new statistics curriculum of 30 to 40 lessons in grade 7 and 8 over a two-year period.
The goal of such a curriculum is that students become statistically literate, in particular they learn to analyze data and communicate about statistical information. To achieve this, students should extend their case-oriented views with aggregate views of data. For describing and predicting aggregate features of data sets, particular statistical key concepts turn out to be indispensable: variability, sampling, data, center, spread, and distribution. The most fundamental key concepts in statistics are variability and uncertainty. Without a sense of the variability of a certain phenomenon (e.g. life span of batteries), there is no reason for students to think of a sample or a distribution either. It is therefore important to choose a problem situation in which students expect variability or acknowledge the uncertainty involved in the context. One suitable problem situation is estimation of large numbers.
The curriculum focuses on the key concepts of data, center, spread, sampling, and distribution (shape) in relation to case-value plots such as value-bar graphs and dot plots. These concepts and diagram types have to become tools in diagrammatic reasoning; for instance, when comparing distributions or when describing stable distribution aspects (for example whilst growing a sample). Though coherent reasoning about these key concepts is favorable, they cannot always be addressed at the same time. We suggest starting with center and spread as the most important distribution aspects (as in Chapter 7), and then address sampling and shape issues. In terms of diagrammatic reasoning, we advocate the following:

1 students make diagrams of data sets and hypothetical situations (diagrammatization);
2 they experiment with diagrams (mental experimentation can be stimulated with what-if questions and physical experimentation is preferably done with a computer tool that allows dynamic interaction with diagrams);
3 they reflect on the diagrams (the teacher and the instructional design are important here).

We would use a computer tool that allows diagrammatization and user-friendly ways of performing genuine data analysis. The Minitools have limited options to diagrammatize (values need to be entered one by one and there is no drawing tool), and when organizing a data set within one Minitool, students can only use one type of plot, such as a dot plot. Tinkerplots (Konold & Miller, 2004), a recently developed construction tool for statistical data analysis, more closely fits our criteria (10.3.4).
Somewhere in the curriculum, students should experience a complete investigative cycle of data analysis from a question, design, sampling, data analysis to communi-

cation of the results and formulation of a new or more precise question (cf. NCTM, 2000; Wild & Pfannkuch, 1999). Every now and then, graphs or messages from the media can be used to foster a critical attitude (De Lange et al., 1993).

From the historical phenomenology, it is clear that we should distinguish three fields in which students can learn different aspects of statistical notions: error theory with repeated measurements, natural phenomena with symmetrical distributions, and social contexts with irregular data. Accordingly, we envision different routes of developing notions of center and spread that probably need to be combined.

The first route roughly follows the historical development of error theory in science. It uses students' notions of a true value when conducting repeated measurements of one item. Konold and Pollatsek (2002) argue that students can thus develop a sense of signal and noise, which can be seen as a conceptual underpinning of understanding average values.[63] This is the route that Petrosino, Lehrer, and Schauble (2003) took in a fourth-grade classroom. We expect students to learn to reason about measurements as being around a true value, that errors on both sides are equally likely, and that large errors are less likely than small errors (H13). The mean, which Dutch seventh graders already know as an algorithm, can be used to estimate the true value. The median can be introduced as the middle-most value (in particular if students think that negative and positive errors are equally likely).

Spread then appears as dispersion from the true value (or the measure of center that estimates that true value). We then let students compare situations with small and large errors to make spread a topic of discussion. The range can be made equal to avoid a conflict between dispersion and range. We anticipate that students come to see the majority of the measurements as a clump of data close together, and that there will always be some low and high values. To guide the reinvention of quartiles, we could ask, "between which two values is about half of the errors?" On the basis of the historical development of quartiles we assume that looking at halves is natural to students (H14). Computer tool options such as four equal groups can become ways of quantifying spread and center if suitable contexts are chosen.

Such a repeated measurement approach could be used for combining science and mathematics lessons (Erickson, 2002). Within the Dutch education system, however, this is difficult to accomplish because Dutch students do not receive instruction in physics until grade 8 and chemistry until grade 9. One drawback of this first route of error theory is that is appears to be at odds with the EDA approach.

The second route of developing notions of center and spread starts from students' knowledge of what is typical and what is not in natural contexts with roughly symmetrical and smooth distributions. As we have seen throughout the teaching experiments, students know what typical weights and heights are for their age. Using such

63. More generally, a distribution, a trend, or a model is a way of capturing a signal, the variation around which is considered noise.

familiar contexts, we can help them express a categorization into three groups of low, typical, and high values, both in a natural and a diagrammatic language. In terms of diagrammatic reasoning it is important that students learn to predicate features of data sets that are represented in diagrams. For example, where is the majority of the data values? These majorities, for instance represented as 'clumps' in dot plots, can function as an initial notion of center. When students are stimulated to more precisely describe where they see these clumps, as a range, these clumps can come to function as an informal measure of spread as well. If two distributions with the same range have the clumps in the same place, but one clump has a larger range than the other, we expect that students will see that the former has a larger spread than the latter (in Figure 10.4 the clumps have different ranges).

As shown in Chapter 7, it is important to provide students with tools that can help them become more precise in their reasoning. For instance, it turns out useful to introduce the notion of range as distinct from spread to avoid confusion in the discourse and enhance student abilities to express what they see in diagrams. If opportunities appear to introduce conventional definitions (e.g., median, mode), we can take advantage of them.

Using such data invention tasks as we used in the battery context, we can stimulate students to express aggregate features of center and spread in diagrams: what would a reliable battery brand with a short life span look like in a diagram? In such activities, reliability is taken as a contextual basis for a notion of spread, and average life span as a basis for center. With the compensation strategy of visually estimating the mean, students can learn that the mean accounts for all data values and how it is influenced by extreme values. Gradually the mean can come to function as measuring the position of a clump (the center of the distribution), for instance when comparing different data sets. This can be done with an activity similar to the speed trap. A disadvantage of the speed trap context is that the speed limit can be a distracting value that evokes reasoning below and above that limit, whereas we try to foster reasoning with centers of distributions as a whole. When comparing different degrees of being spread out, students can use computer tool options such as making their own groups, equal group size, and four equal groups. It should be emphasized that it makes sense to have a convention of using a standard way, for instance four equal groups (quartiles). As we argue in Chapter 7, students should be stimulated to describe how data are spread out. Doing so, they often describe how data are distributed.

From early lessons onwards, attention should be paid to sampling. In the beginning, also in grade 7, this could be done by letting students collect data of something simple (such as reaction time or car color). Another way of addressing sampling is talking through the process of data creation. A more explicit way of addressing sampling is asking students to design a method of testing something. Next, an activity similar to growing samples can be done to stimulate diagrammatic reasoning about distribution aspects (as in Sections 9.4 and 9.5). The advantage of growing a sample is that

it starts from a sample size that students initially find reasonable and ends with the shape of a population distribution. Students expect such features as the mean to be stable after some time, and some may even consider the shape to stabilize. The situation of comparing predictions and real data sets supports students to reflect in aggregate terms, because comparing individual cases is clearly not very helpful.
When reasoning about shapes they propose themselves, in addition to skewed shapes we think need to be discussed, students can use the statistical notions they know and show the understanding they have of shape. As in Chapter 9, we expect students to acknowledge that there are only a few low and high values in most unimodal distributions. This is represented as low horizontal parts in a continuous shape, and the large group of average values is represented as a high flat part in the shape; and of course, there are values in between that cause the slopes of the shape.
In higher grades, we can ask students to describe new distribution shapes, such as normal, bimodal, uniform, and skewed, all represented in dot plots. The data sets should be large enough for students to recognize the shapes and come to see them as signals in noise. Next, students learn to operate with these shapes as objects, for instance mentally shifting them along the axis or predicting shapes of new situations. At some point, students can handle social science contexts in which irregular data are not so easy to model with distributions.
We anticipate that a curriculum, as described here, leads to statistical understanding that forms the basis for a more formal introduction of the normal distribution in higher grades. More generally, we expect that the approach promoted here contributes to statistical literacy (Gal, 2002).

10.5 Recommendations for teaching, design, and research

10.5.1 Recommendations for teaching and instructional design

Statistics chapters in most mathematical textbooks introduce students to the important statistical notions and graphs in what we call a ‘topic-topic-topic approach’. The rationale of such an approach is probably that once students have mastered those statistical notions and graphs, they will have learned statistics or at least be prepared to carry out statistical analyses. In Chapter 2, we criticize this approach of introducing the theory before the application: research in statistics education shows that this mostly leads to disappointing results. The mean, for instance, is more than an algorithm performed on data values. As shown in several chapters, the mean has many qualitative and quantitative aspects. To know the algorithm is certainly not to know when to use the mean or to know how to use it as a group descriptor or a representative value. Similarly, it is not difficult for middle school students to read values from a histogram, but it is demanding to interpret a histogram as a diagram or a symbol of a frequency distribution. Therefore, researchers in statistics education have

been in search of ways to engage students in genuine data analysis while offering opportunities to develop statistical notions and graphs as meaningful tools. The present research is within this tradition.

Though the present design research focused on instructional design and students' learning, we noticed that the influence of the teacher could hardly be overestimated. In previous chapters, we mentioned a few issues for the instruction theory in which the role of the teacher is pivotal.

First, *talking through the process of data creation* helps to bring contexts to life, so that students will know what the data values stand for. This process takes time, which implies that one context per lesson is to be preferred over many contexts per lesson (cf. Van den Boer, 2003). One of the core questions is: how do you think the data were collected? As shown in Chapter 9, it can be very informative to let students design an investigation.

Second, the teacher as well as the designer can stimulate students' *roles as data analysts* to ensure that students do not just solve the problem at issue, but also think about ways to communicate the results clearly to others such as decision makers and discuss the statistics at issue. This includes reflection on the conveying power of diagrams (cf. DiSessa, 2002).

Third, good *practices and norms* do not emerge automatically. We have noticed that students do not automatically understand that they should base their arguments on the available data. Students are not always used to listening to each other and asking questions if they do not understand what the teacher or their classmates are saying (cf. Yackel & Cobb, 1996). As a matter integral to the learning process, it is necessary to come to taken-as-shared topics of discussion. Additionally, a what-if attitude ("what would a diagram look like if...?") is only developed if a teacher establishes a practice of asking what-if questions.

Fourth, we observed that class discussions were difficult to organize in the computer lab, because students are easily distracted by what is visible on the screens. A practical suggestion is to have students switch off the monitors when discussing in the computer lab. For creating a common object of attention, we prefer *discussion in the regular classroom* using slides of screen shots or, even better, a computer projector.

From the semiotic analyses, it follows that the steps of diagrammatic reasoning are key elements in learning statistics. Teachers are therefore recommended to stimulate students to diagrammatize aggregate features, to experiment with diagrams in software and in the mind (by establishing a habit of asking 'what-if' questions), and to reflect. The reflection phase includes students precisely describing what they see in diagrams (e.g. clump, majority, bump, but also the shape of the bump) and where they see it (from which value to which value). This works into two directions: students learn to express their thoughts as part of their conceptual development, and teachers can assess students' statistical ideas. The insights acquired are a prerequisite for supporting students in their conceptual development.

Experts such as teachers and researchers may easily interpret students' explanations in a more advanced way than students actually think. For instance, when students first used the term 'bump', we thought they referred to the whole shape, but from the retrospective analyses we concluded that they initially referred to the 'majority' of the data or the top part of the whole shape only. Another tricky notion is 'most'. If students say 'most' they can refer to many different things such as the highest values, the most frequent values, or the largest group. At an early stage, we often unconsciously chose the most plausible interpretation, but a closer look sometimes proved us wrong.

The semiotic analyses show how intimately conceptual development is linked to the development of a suitable natural and diagrammatic language. From a semiotic point of view it is therefore important that the classroom interaction focuses on clear topics of common attention, which are visible in a diagram and described in the natural language, because this is the way objects can be formed (in Peirce's terms, what is talked or thought about is an object). Discussing such objects or topics of common attention provides opportunities for hypostatic abstraction and thus supports students' conceptual development.

Throughout the chapters, we have presented a few design heuristics for instructional design in statistics education. The main goal, in our view, is for students to learn to think in aggregate terms about a data set as a whole. To avoid a case-oriented view, it can sometimes help to "stay away from data," as we refer to in Chapter 6. Related to this, it can help to "ask about forests instead of trees," for instance by asking questions about hypothetical situations (e.g. weight diagram of a group of older students, the shape of a larger sample). Furthermore, comparing multiple representations can, but need not, support students in using conceptual tools. For instance, when asked about the common ground of two different representations of the same data set, students may use and develop statistical notions that help in describing the common feature (such as spread or bump). However, we have also observed that students easily resort to comparing superficial features of the representations (cf. Seeger, 1998, Van Someren et al., 1998). In that case, students do not interpret the signs as diagrams or symbols, but compare iconic or indexical elements of the signs.

What are the consequences of the semiotic analyses for instructional design? Just as in the recommendations for teaching, instructional designers are recommended to stimulate the steps of diagrammatic reasoning and create opportunities for predication and hypostatic abstraction. Our experience is that conveying the core ideas of a hypothetical learning trajectory to teachers is far from trivial, which means that instructional designers are recommended to think of ways of supporting teachers helping their students.

For criteria for selecting statistics software, we refer to Section 10.3.4.

10.5.2 Recommendations for future research

Statistics education research is a relatively young discipline, especially the domain of using computer tools dedicated to *learning* statistical data analysis as opposed to *performing* data analysis. In this section we present a few research challenges for statistics education in which using computer tools is integral to learning statistical data analysis.

As Noss and Hoyles (1996) write, "new technologies – *all* technologies – inevitably alter how knowledge is constructed and what it means to any individual" (p. 106). This raises the epistemological question of how using statistics tools changes students' learning and the type of knowledge they acquire. This influence may be rather drastic: from statisticians we have heard that the software they use is so important for their way of working and thinking that they often characterize themselves according to the software they use (e.g., "I am a Minitab statistician").

Computer tools have some evident advantages: students can dynamically interact with large data sets and different graphical representations in a way that is impossible by hand. Using such tools may shift the focus from calculating means and drawing histograms of small data sets to exploring large data sets with multiple representations and ready available means and medians. Furthermore, computer tools such as the Minitools and Tinkerplots offer ways of grouping data into four equal groups and equal intervals that are laborious by hand, but prepare insight into distributions as seen in box plots and histograms.

However, computer tools can easily be almost independent worlds with their own rules and peculiarities. In that case, the software itself needs to be learned before it can effectively mediate between the learner and what is to be learned. This would not be problematic if learning software were part of the curriculum but, in practice, people expect learning with computers to lead to similar knowledge in faster and better ways. It would be more realistic to expect learning and the acquired knowledge to change; how is a topic of further research.

We see at least two theoretical frameworks that can help to gain a deeper understanding of the influence of using IT tools on student learning. The first is the instrumental approach (Artigue, 2002; Drijvers, 2003), which theorizes on how artefacts such as IT tools become instruments for solving mathematical problems in students' hands. As Drijvers has shown for learning algebra with computer algebra systems, the development of 'instrumentation schemes' always has a technical and a conceptual side, which are interwoven. It is certainly not the case that having an IT tool that does the laborious computations allows students to focus on the conceptual side of solving problems (Artigue, 2002). Technical peculiarities of the tool can hinder students, but sometimes can also be used to improve students' conceptual understanding (see Drijvers, 2003, for examples). It is likely that similar and different observations can be found for the instrumentation process of using statistics tools for solving statistical problems. One difference is that the educational statistics software we report on

is specially designed for middle school students and has fewer technical hurdles.
A second theoretical framework that we consider potentially fruitful is a semiotics of IT tools, which still has to be worked out. When semiotics came to birth, signs were static. Though Peirce's notion of interpretant offers the opportunity to stress the dynamic aspects of sign activity, the interpretant is only the interpreter's response, not the reaction of a computer tool to the interpreter's actions. A computer tool reacts according to hidden rules and is therefore not transparent to all users. In our view, an elaborated sign notion is needed that takes the dynamic interaction with computer tools into account.
So far, we have focused on the influence of using IT tools on learning, but it is evident that drastic changes in learning demands other ways of teaching, assessing, and designing.

Teaching. In the present research we have focused on students' learning and have kept the role of the teacher in the background. Partially we could do so because we worked with experienced teachers who were well acquainted with the RME philosophy. However, in a follow-up teaching experiment we would certainly focus more on the teacher's role. One reason is that the success of particular instructional activities depends on the teacher. Another reason is that the approach we have taken in this study is not easy for teachers. They have to lead class discussions on students' reasoning about a variety of diagrams and arguments. The more options a computer tool offers, the more difficult it seems to be for the teacher. Particularly if the teacher wants to let students benefit from their own ideas, she or he needs to see the potential of students' often imprecise formulations. Furthermore, this approach requires a good understanding of data analysis techniques and of students' learning of those techniques. This implies the need for research into the professionalization of teachers in statistical data analysis, especially if software is used. In particular, we see the need of investigating how teachers can grasp the core ideas of an HLT.

Assessment. Nor is students' learning in the approach of this study easy to assess. We have used a number of assessment tasks, some of which were more informative than others, but most of them were rather laborious to score (e.g. question 2 from the test in class 1B). Such questions are unlikely to be used in national tests. Therefore, there is a need of assessment items that do assess what students learned and that might be used in large-scale tests (cf. Konold & Khalil, 2003).

Design. The hypotheses mentioned in Section 10.3.4 lead to the question of what the relative merits are of route and landscape-type tools for learning data analysis. In this thesis we have formulated a few questions about the affordances and constraints of particular representations such as value-bar graphs and dot plots. Do students, within settings as created in the present research, more readily see spread in a dot plot than in a value-bar graph? Can students more easily estimate the mean in a value-bar graph than in a dot plot or histogram? Does it make a difference for learning about

the median in a value-bar graph whether the bars are vertical or horizontal? The answers to such questions are necessary to inform future designs, of instructional materials including software. Nonetheless, we were unable to answer those questions within the methodology chosen. Nor do we think that purely quantitative comparative research can help to answer those questions, because those diagrams gain their meanings within communities of learners (Meira, 1998). A combination of design research that creates the right conditions, and comparative research that isolates specific aspects may yield empirically grounded answers to such questions. Such a set-up requires a larger team than we had.

Finally, we need to gain more insight into which types of statistics education lead to statistical literacy so that citizens of the future knowledge society will wisely use its most important resource: information.

Appendix

The series of the most important conjectures that were generated and confirmed during the retrospective analyses in classes 1B and 2B are the following.

C1. Students divide imaginary data sets into three groups of low, 'average', and high values.

C2. Students either characterize spread as range or look very locally. We call the first view a range view of data and the second a density view ("here the dots are close to each other and there they are spread out"). There are no examples of views of spread as dispersion from a measure of center.

C3. Students are inclined to think of small samples when first asked about how one could test something (batteries, weight).

C4. Students do not expect variation in industrial contexts such as the battery life span context.

C5. What-if questions work well for letting students think of aggregate features of a graph or a situation. What would a weight graph of older students look like? What would the graph look like if a larger sample was taken? What would a larger sample of a good battery brand look like?

C6. If students have to draw their own graphs, they often draw vertical value-bar graphs although Minitool 1 only offers horizontal bars.

C7. Students' notions of spread, distribution, and density are not yet distinguished. When explaining how data are spread out, they often describe the distribution or the density in some area.

C8. Students often mistake the median for the midrange.

C9. Even when students see a large sample of a particular distribution, they often do not see the shape we see in it (lesson 8 of 2B).

References

ACE (1991). *A national statement on mathematics for Australian schools*. Carlton, Vic, Australia: Curriculum Corporation.

Ainley, J. (2000). Transparency in graphs and graphing tasks. An iterative design process. *Journal of Mathematical Behavior, 19*, 365-384.

Ainley, J., Nardi, E., & Pratt, D. (2000). The construction of meanings for trend in active graphing. *International Journal of Computers for Mathematical Learning, 5*, 85-114.

Ainsworth, S., Bibby, P., & Wood, D. (2002). Examining the effects of different multiple representational systems in learning primary mathematics. *Journal of the Learning Sciences, 11*, 25-61.

Anderson, J. R. (1995). *Cognitive psychology and its implications* (Fourth ed.). New York: W.H. Freeman and Company.

Anderson, M., Sáenz-Ludlow, A., Zellweger, S., & Cifarelli, V. V. (Eds.). (2002). *Educational perspectives on mathematics as semiosis: From thinking to interpreting to knowing*. Ottawa, Canada: Legas Publishing.

Aristotle (1994). *Nicomachean ethics*. Cambridge, MA: Harvard University Press.

Arnould, J., & Maclachlan, D. (1872). *On the law of marine insurance* (Fourth Ed.). London: Stevens.

Artigue, M. (2002). Learning mathematics in a CAS environment: The genesis of a reflection about instrumentation and the dialectics between technical and conceptual work. *International Journal of Computers for Mathematical Learning, 7*, 245-274.

Ashburner, W. (1909). *Nomos Rhodioon nautikos; The Rhodian sea-law*. Oxford: Clarendon Press.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2002). *The resilience of overgeneralization of knowledge about data representations*. Presented at the American Education Research Association conference.

Bakker, A. (1999). Statistisch redeneren; Uitstapje naar de cognitieve psychologie [Statistical reasoning from a cognitive psychology perspective]. *Nieuwe Wiskrant, 19*(2), 42-46.

Bakker, A. (2002). Route-type and landscape-type software for learning statistical data analysis. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference of Teaching Statistics [CD-ROM]*. Voorburg, the Netherlands: International Statistical Institute.

Bakker, A. (2003). The early history of statistics and implications for education. *Journal of Statistics Education, 11*(1). www.amstat.org/publications/jse/v11n1/bakker.html.

Bakker, A., & Gravemeijer, K. P. E. (2003). Planning for problem solving in statistics. In R. Charles & H. L. Schoen (Eds.), *Teaching mathematics through problem solving: Grades 6 - 12* (pp. 105-117). Reston, VA: National Council of Teachers of Mathematics.

Barab, S. A., & Kirshner, D. (2001). Guest editor's introduction: Rethinking methodology in the learning sciences. *Journal of the Learning Sciences, 10*, 5-15.

Beniger, J. R., & Robyn, D. L. (1978). Quantitative graphics in statistics: A brief history. *The American Statistician, 32*(1), 1-11.

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics, 45*, 35-65.

Ben-Zvi, D., & Friedlander, A. (1997). Statistical thinking in a technological environment. In J. B. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning of statistics*. Voorburg, the Netherlands: International Statistical Institute.

Ben-Zvi, D., & Garfield, J. (Eds.). (in press). *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Bereiter, C. (1985). Towards a solution of the learning paradox. *Review of Educational Research, 55*, 201-226.

Bereiter, C. (2002). Design research for sustained innovation. *Cognitive Studies, Bulletin of the Japanese Cognitive Science Society, 9*, 321-327.

Bernoulli, D. (1982). *Die Werke von Daniel Bernoulli [The works of Daniel Bernoulli]* (B. L. van der Waerden, Ed.). Basel, Switzerland: Birkhäuser Verlag.

Bethlehem, J., & De Gooijer, J. (2000). *Data-analyse [data analysis]*. Amsterdam, the Netherlands: University of Amsterdam.

Bethlehem, J., & De Ree, S. J. M. (1999). *100 jaar CBS: Van populatie naar steekproef [100 years CBS: From population to sample]* (No. 9901). Voorburg: Centraal Bureau voor de Statistiek.

Biehler, R. (1982). *Explorative Datenanalyse - eine Untersuchung aus der Perspektive einer deskriptiv-empirischen Wissenschaftstheorie [Explorative data analysis - an investigation from the perspective of a descriptive-empirical scientific theory]*. Bielefeld: Universität Bielefeld.

Biehler, R. (1994). *Probabilistic thinking, statistical reasoning, and the search for causes: Do we need a probabilistic revolution after we have taught data analysis?* Paper presented at the Fourth International Conference on Teaching Statistics, Marrakech.

Biehler, R. (1997). Software for learning and for doing statistics. *International Statistical Review, 65*, 167-189.

Biehler, R., & Steinbring, H. (1991). Entdeckende Statistik, Stengel-und-Blätter, Boxplots: Konzepte, Begründungen und Erfahrungen eines Unterrichtsversuches [Statistics by discovery, stem-and-leaf, boxplots: Basic conceptions, pedagogical rationale, and experiences from a teaching experiment]. *Der Mathematikunterricht, 37*(6), 5-32.

Bissell, D. (1996). Statisticians have a word for it. *Teaching Statistics, 18*, 87-89.

Boswinkel, N., Miehaus, J., Gravemeijer, K. P. E., Middleton, J. A., Spence, M. S., Burrill, G., et al. (1997). *Picturing numbers*. Chicago: Encyclopaedia Britannica Educational Corporation.

Box, G. P. (1999). Statistics as a catalyst to learning by scientific method part II–a discussion. *Journal of Quality Technology, 31*(1), 16-29.

Boyer, C. B. (1991). *A history of mathematics* (Revised edition by U.C. Merzbach ed.). New York: John Wiley & Sons, Inc.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2*, 141-178.

Cobb, G. W. (1997). Mere literacy is not enough. In L. A. Steen (Ed.), *Why numbers count: Quantitative literacy for tomorrow's America* (pp. 75-90). New York: College Enterance Examination Board.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104*, 801-824.

Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning, 1*, 5-43.

Cobb, P. (2000). From representations to symbolizing: Introductory comments on semiotics and mathematical learning. In P. Cobb, E. Yackel & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms; Perspectives on discourse, tools, and instructional design* (pp. 17-36). Mahway, NJ: Lawrence Erlbaum Associates.

Cobb, P. (2002). Modeling, symbolizing, and tool use in statistical data analysis. In K. P. E. Gravemeijer, R. Lehrer, B. van Oers & L. Verschaffel (Eds.), *Symbolizing, modeling and tool use in mathematics education* (pp. 171-196). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Cobb, P., Confrey, J., diSessa, A. A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*, 9-13.

Cobb, P., Gravemeijer, K. P. E., Bowers, J., & Doorman, M. (1997). *Statistical Minitools* [applets and applications]. Nashville & Utrecht: Vanderbilt University, TN & Freudenthal Institute, Utrecht University. See www.wisweb.nl

Cobb, P., Gravemeijer, K. P. E., Yackel, E., McClain, K., & Whitenack, J. (1997). Mathematizing and symbolizing: The emergence of chains of signification in one first-grade class-

room. In D. Kirshner & J. A. Whitson (Eds.), *Situated cognition theory: Social, semiotic, and neurological perspectives* (pp. 151-233). Hillsdale, NJ: Erlbaum.
Cobb, P., & Hodge, L. L. (2002a). Learning, identity, and statistical data analysis. In B. Phillips (Ed.), *Developing a Statistically Literate Society; Proceedings of the Sixth International Conference of Teaching Statistics [CD-ROM]*. Voorburg, the Netherlands: International Statistics Institute.
Cobb, P., & Hodge, L. L. (2002b). A relational perspective on issues of cultural diversity and equity as they play out in the mathematics classroom. *Mathematical Thinking and Learning, 4*, 249-284.
Cobb, P., & McClain, K. (2002). Supporting students' learning of significant mathematical ideas. In G. Well & G. Claxton (Eds.), *Learning for life in the 21st century: Sociocultural perspectives on the future of education* (pp. 154-166). Oxford, UK: Blackwell.
Cobb, P., & McClain, K. (in press). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecht, the Netherlands: Kluwer Academic Publisher.
Cobb, P., McClain, K., & Gravemeijer, K. P. E. (2003). Learning about statistical covariation. *Cognition and Instruction, 21*, 1-78.
Cobb, P., & Tzou, C. (2000). *Learning about data creation.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
Cobb, P., & Whitenack, J. (1996). A method for conducting longitudinal analyses of classroom videorecordings and transcripts. *Educational Studies in Mathematics, 30*, 213-228.
Cobb, P., Yackel, E., & McClain, K. (Eds.). (2000). *Symbolizing and communicating in mathematics classrooms; Perspectivies on discourse, tools, and instructional design*. Mahwah, NJ: Lawrence Erlbaum Associates.
Cohen, I. B. (1984). Florence Nightingale. *Scientific American, March*, 98-108.
Cole, M. (1996). *Cultural psychology: A once and future discipline*. London: Belknap Press.
Collins, A. (1992). Toward a design science of education. In E. Scanlon & T. O'Shea (Eds.), *New directions in educational technology* (pp.15-22). New York: Springer Verlag.
Condorcet (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: l'Imprimerie Royale. [Reprint: 1972 by Chelsea Publishing Company: New York]
Cortina, J. L. (2002). Developing instructional conjectures about how to support students' understanding of the arithmetic mean as a ration. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference of Teaching Statistics [CD-ROM]*. Voorburg, the Netherlands: International Statistical Institute.
Cortina, J. L., Saldanha, L., & Thompson, P. W. (1999). Multiplicative conceptions of the arithmetic mean. In F. Hitt & M. Santos (Eds.), *Proceedings of the Twenty First Meeting of the North American Chapter of the International Group of the Psychology of Mathematics Education* (Vol. 2, pp. 466-472). Cuernacava, Morelos, Mexico: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.
Cournot, A. A. (1843). *Exposition de la théorie des chances et des probabilités [Treatise of the theory of chances and probabilities]*. Paris: Librairie philosophique J. Vrin.
David, H. A. (1995). First (?) occurences of common terms in mathematical statistics. *The American Statistician, 49*, 121-133.
David, H. A. (1998a). Early sample measures of variability. *Statistical Science, 13*, 368-377.
David, H. A. (1998b). First (?) occurences of common terms in probability and statistics--A second list, with corrections. *The American Statistician, 52*, 36-40.
De Corte, E. (2000). Marrying theory building and the improvement of school practice: A permanent challenge for instructional psychology. *Learning and Instruction, 10*, 249-266.
De Knecht-van Eekelen, A., & Stamhuis, I. H. (1992). De met cijfers bedekte negentiende eeuw. Toepassing van statistiek en waarschijnlijkheidsrekening in Nederland en Vlaanderen tussen 1840 en 1920. *Gewina, 15*, 137-139.

De Lange, J. (1987). *Mathematics, insight and meaning*. Utrecht, the Netherlands: OW&OC, Rijksuniversiteit Utrecht.

De Lange, J., Burrill, G., Romberg, T., & Van Reeuwijk, M. (1993). *Learning and testing Mathematics in Context: The case: data visualization*. Madison, WI: University of Wisconsin, National Center for Research in Mathematical Sciences Education.

De Mast, F. (1998). *Rondneuzen in de statistieke historie [Nose around in the history of statistics] (preprint version)*. Voorburg, the Netherlands: Centraal Bureau voor de Statistiek.

De Mast, J. (2002). *Quality improvement from the viewpoint of statistical method*. Amsterdam: University of Amsterdam.

Dekker, R., & Elshout-Mohr, M. (1998). A process model for interaction and mathematical level raising. *Educational Studies in Mathematics, 35*, 303-314.

Descamps, K., Janssens, D., & Vanlangendonck, B. (2001). Statistiek op de werkvloer [Statistics in the workplace]. *Nieuwe Wiskrant, 20*(1), 4-8.

Dictionary of Art (1996). J. S. Turner (Ed.). New York: Grove.

Dijksterhuis, E. J. (1950/1986). *The mechanization of the world picture: Pythagoras to Newton* (C. Dikshoorn, Trans.). Princeton, NJ: Princeton University Press.

Dijksterhuis, E. J. (1990). *Clio's stiefkind [Clio's stepchild]*. Amsterdam: Bert Bakker.

DiSessa, A. A. (2002). Students' criteria for representational adequacy. In K. P. E. Gravemeijer, R. Lehrer, B. van Oers & L. Verschaffel (Eds.), *Symbolizing, modeling and tool use in mathematics education* (pp. 105-130). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Doerr, H. M., & Zangor, R. (2000). Creating meaning for and with the graphing calculator. *Educational Studies in Mathematics, 41*, 143 - 163.

Doerr, H. M., & English, L. D. (2003). A modeling perspective on students' mathematical reasoning about data. *Journal for Research in Mathematics Education, 34*, 110-136.

Doorman, L. M. (in press). *Design research on teaching and learning the integrated principles of calculus and kinematics with ICT*. Utrecht, the Netherlands: CD Beta Press.

Dörfler, W. (1993). Computer use and views of the mind. In C. Keitel & K. Ruthven (Eds.), *Learning from computers: Mathematics education and technology* (pp. 159-186). Berlin: Springer Verlag.

Dörfler, W. (2000). Means for meaning. In P. Cobb, E. Yackel & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms; Perspectivies on discourse, tools, and instructional design* (pp. 99-131). Mahwah, NJ: Lawrence Erlbaum Associates.

Dörfler, W. (2003). Diagrams as means and objects of mathematical reasoning. In *Developments in mathematics education in German-speaking countries. Selected papers from the annual conference on didactics of mathematics*.

Drijvers, P. (2003). *Learning algebra in a computer algebra environment: Design research on the understanding of the concept of parameter*. Utrecht, the Netherlands: CD Beta Press.

Dubinsky, E. (1991). Reflective abstraction in advanced mathematical thinking. In D. Tall (Ed.), *Advanced mathematical thinking* (pp. 95-123). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Ducrot, O., & Todorov, T. (1983). *Encyclopedic dictionary of the sciences of language* (C. Porter, Trans.). Baltimore: Johns Hopkins University Press.

Eco, U. (1984). *Semiotics and the philosophy of language*. London: MacMillan Press.

Edelson, D. C. (2002). Design research: What we learn when we engage in design. *Journal of the Learning Sciences, 11*, 105-121.

Editors of JMB (2001). A garland of teaching experiments. *Journal of Mathematical Behavior, 20*, 263-266.

Eisele, C. (1976). Peirce's philosophy of mathematical education. In R. M. Martin (Ed.), *Studies in the scientific and mathematical philosophy of Charles S. Peirce. Essays by Carolyn Eisele*. The Hague, the Netherlands: Mouton Publishers.

Eisenhart, C. (1974). *The development of the concept of the best mean of a set of measurements from antiquity to the present day. 1971 ASA Presidential Address*. Unpublished manu-

script.
Eisenhart, C. (1977). Boscovich and the combination of observations. In M. G. Kendall & R. L. Plackett (Eds.), *Studies in the history of statistics and probability* (Vol. 2). London: Charles Griffin & Company Limited.
Engeström, Y. (1987). *Learning, working and imagining: Twelve studies in activity theory*. Helsinki: Orienta-Konsultit.
Erickson, T. (2002). Technology, Statistics, and Subtleties of Measurement: Bridging the Gap Between Science and Mathematics. In B. Phillips (Ed.), *Developing a Statistically Literate Society. Proceedings of the Sixth International Conference on Teaching Statistics*. Cape Town, South Africa.
Erkens, G. (2001). *MEPA: Multiple Episode Protocol Analysis* (Version 4.6.2). Utrecht: Utrecht University.
ESS (1981). *Encyclopedia of statistical sciences* (Eds. Kotz, S. & Johnson, N.L.). New York: Wiley & Sons.
Euclid (1956). *The thirteen books of The Elements. Translation with introduction and commentary by Sir Thomas L. Heath* (T. H. Heath, Trans.). New York: Dover.
Fauvel, J., & Van Maanen, J. (Eds.). (2000). *History in mathematics education*. Dordrecht, the Netherlands: Kluwer Academic Publishers.
Feldman, A., Konold, C., & Coulter, B. (2000). *Network science, a decade later; The Internet and classroom learning*. Mahwah, NJ: Lawrence Erlbaum Associates.
Feldman, J., Lagneau, G., & Matalon, B. (Eds.). (1991). *Moyenne, milieu, centre: histoire et usages [Mean, middle, center: history and uses]*. Paris: Ecole des Hautes études en Sciences Sociales.
Fosnot, C. T., & Dolk, M. (2001). *Young mathematicians at work. Constructing number sense, addition, and subtraction*. Portsmouth, NH: Heinemann.
Frege, F. L. G. (1962). *Funktion, Begriff, Bedeutung: Fünf logische Studien*: Göttingen.
Frege, F. L. G. (1895/1976). *Wissenschaftlichter Briefwechsel [Scientific correspondence]* (Gabriel, G., Ed.) (Vol. 2). Hamburg, Germany: Meiner.
Freudenthal, H. (1966a). *De eerste ontmoeting tussen de wiskunde en de sociale wetenschappen [The first meeting of mathematics with the social sciences]*. Brussel: Paleis der Academiën.
Freudenthal, H. (1966b). *Waarschijnlijkheid en Statistiek [Probability and Statistics]*. Haarlem: De Erven F. Bohn.
Freudenthal, H. (1973). *Mathematics as an educational task*. Dordrecht, the Netherlands: Reidel.
Freudenthal, H. (1974). Waarschijnlijkheid en statistiek op school [Probability and statistics at school]. *Euclides, 49*, 245-246.
Freudenthal, H. (1983a). *Didactical phenomenology of mathematical structures*. Dordrecht, the Netherlands: Reidel.
Freudenthal, H. (1983b). The implicit philosophy of mathematics: History and education. In *Proceedings of the International Congress of Mathematicians* (pp. 1695-1709). Warsaw.
Freudenthal, H. (1991). *Revisiting mathematics education: China lectures*. Dordrecht, the Netherlands: Kluwer Academic Publishers.
Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*, 124-158.
Gal, I. (2002). Adult's statistical literacy: meanings, components, responsibilities. *International Statistical Review, 70*, 1-51.
Gal, I., & Garfield, J. B. (Eds.). (1997). *The assessment challenge in statistics education*. Amsterdam: IOS Press.
Galton, F. (1869). *Hereditary genius: an inquiry into its laws and consequences*. London: Macmillan.
Galton, F. (1875). Statistics by intercomparison, with remarks on the law of frequency of error.

Philosophical Magazine, 49(4), 33-46.
Galton, F. (1883). *Inquiries into human faculty and its development*. London: Dent and co.
Galton, F. (1889a). *Narrative of an explorer in tropical south Africa being an account of a visit to Damaraland in 1851*. London: Ward, Lock and co.
Galton, F. (1889b). *Natural inheritance*. London: Macmillan.
Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in statistics: Implication for research. *Journal for Research in Mathematics Education, 19*, 44-63.
Gillies, D. (Ed.). (1992). Revolutions in mathematics. Oxford, UK: Clarendon Press.
Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory; Strategies for qualitative research*. Chicago: Aldine Publishing Company.
Godard, R., & Crépel, P. (1999). An historical study of the median. In *Proceedings of the CSHPM* (pp. 207-218).
Goldin, G. A., & Janvier, C. (1998). Representations and the Psychology of Mathematics Education. *Journal of Mathematical Behavior, 17*, 1-4.
Gravemeijer, K. P. E. (1994). *Developing realistic mathematics education*. Utrecht: CD Bèta Press.
Gravemeijer, K. P. E. (1998a). Developmental research as a research method. In J. Kilpatrick & A. Sierpinska (Eds.), *Mathematics Education as a Research Domain: A Search for Identity (An ICMI Study)* (Vol. 2, pp. 277-295). Dordrecht: Kluwer Academic Publishers.
Gravemeijer, K. P. E. (1998b). Symboliseren en modelleren als wiskundige activiteit [Symbolizing and modeling as mathematical activities]. *Tijdschrift voor nascholing en onderzoek van het reken-wiskundeonderwijs, 16*(2/3), 11-18.
Gravemeijer, K. P. E. (1999a). How emergent models may foster the constitution of formal mathematics. *Mathematical Thinking and Learning, 1*, 155-177.
Gravemeijer, K. P. E. (1999b). *ICT in dienst van het heruitvinden van statistische concepten en representaties [Information and communication technology for reinventing statistical concepts and representations]*. Paper presented at ORD (OnderwijsResearchDagen) 1999. Nijmegen, the Netherlands.
Gravemeijer, K. P. E. (1999c). *An instructional sequence of analysing univariate data sets.* Paper presented at the AERA 1999, Montréal, Canada.
Gravemeijer, K. P. E. (2000a). *Didactical phenomenology as a basis for instructional design; data analysis as an example.* Unpublished manuscript, Utrecht, the Netherlands.
Gravemeijer, K. P. E. (2000b). *A rationale for an instructional sequence for analyzing one and two-dimensional data sets*. Unpublished manuscript, Utrecht, the Netherlands.
Gravemeijer, K. P. E. (2001). *Developmental research, a course in elementary data analysis as an example.* Paper presented at The Netherlands and Taiwan Conference on Common Sense in Mathematics Education. Taipei, Taiwan, November 2001.
Gravemeijer, K. P. E. (2002). Emergent modeling as the basis for an instructional sequence on data analysis. In B. Phillips (Ed.), *Developing a Statistically Literate Society; Proceedings of the Sixth International Conference of Teaching Statistics [CD-ROM]*. Voorburg, the Netherlands: International Statistics Institute.
Gravemeijer, K. P. E., & Cobb, P. (2001). *Designing classroom-learning environments that support mathematical learning.* Presented at AERA, Seattle.
Gravemeijer, K. P. E., Cobb, P., Bowers, J., & Whitenack, J. (2000). Symbolizing, modeling, and instructional design. In P. Cobb, E. Yackel & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms: Perspectives on discourse, tools, and instructional design* (pp. 225 - 273). Mahwah, NJ: Lawrence Erlbaum Associates.
Gravemeijer, K. P. E., Lehrer, R., Van Oers, B., & Verschaffel, L. (Eds.). (2002). *Symbolizing, modeling and tool use in mathematics education*. Dordrecht, the Netherlands: Kluwer Academic Publishers.
Gulikers, I., & Blom, K. (2001). 'A historical angle', a survey of recent literature on the use and value of history in geometrical education. *Educational Studies in Mathematics, 47*, 223-258.

Hacking, I. (1975). *The emergence of probability. A philosophical study of early ideas about probability, induction and statistical inference*. London: Cambridge University Press.
Hacking, I. (1990). *The taming of chance*. Cambridge: Cambridge University Press.
Hald, A. (1990). *A history of probability and statistics and their applications before 1750*. New York: John Wiley & Sons.
Hall, M. (2000). *Bridging the gap between everyday and classroom mathematics: An investigation of two teacher's intentional use of semiotic chains.* Unpublished Ph.D. manuscript, The Florida State University.
Hancock, C. (1995). Tabletop [data analysis software for middle school]. Boston, MA: TERC.
Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic enquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 27*, 337-364.
Hardiman, P. T., Well, A. D., & Pollatsek, A. (1984). Usefulness of balance model in understanding the mean. *Journal of Educational Psychology, 76*(5), 792-801.
Harter, H. L. (1977). *A chronological annotated bibliography on order statistics* (Vol. 1): US Air Force.
Heath, T. H. (1981). *A history of Greek mathematics*. New York: Dover.
Hilts, V. L. (1975). *A guide to Francis Galton's English men of science.* Philadelphia: American Philosophical Society.
Hoek, D. J., Seegers, G., & Gravemeijer, K. P. E. (in press). The use of a graphic calculator to solve problems during small group work in a context of vocational education. *Educational Studies in Mathematics*.
Hoffmann, M. (2001). Skizze einer semiotischen Theorie des Lernens [Sketch of a semiotic learning theory]. *Journal für Mathematik-Didaktik, 22*(3/4), 231-251.
Hoffmann, M. H. G. (2001). The 1903 Classification of Triadic Sign-Relations. In J. Queiroz (Ed.), *Digital Encyclopedia of Charles S. Peirce* (Online http://www.digitalpeirce.org/hoffmann/sighof.htm).
Hoffmann, M. H. G. (2002). Peirce's "Diagrammatic Reasoning" as a Solution of the Learning Paradox. In G. Debrock (Ed.), *Process Pragmatism: Essays on a Quiet Philosophical Revolution* (pp. 147-174). Amsterdam: Rodopi Press.
Hoffmann, M. H. G. (2003a). *Erkenntnisentwickelung. Ein semiotisch-pragmatischer Ansatz (Habilitationsschrift) [Development of knowledge. A semiotic-pragmatic treatise]*. Dresden, Germany: Philosophische Fakultät der Technischen Universität.
Hoffmann, M. H. G. (2003b). Semiotik als Analyse-Instrument [Semiotics as an instrument of analysis]. In M. H. G. Hoffmann (Ed.), *Mathematik verstehen - Semiotische Perspektiven [Understanding mathematics - semiotic perspectives]* (pp. 34-77). Hildesheim, Germany: Franzbecker.
Hopkins, M. (1859). *A handbook of average* (Second ed.). London: Smith, Elder, and co.
Hoyles, C., & Noss, R. (2003). What can digital technologies take from and bring to research in mathematics education? In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick & F. K. S. Leung (Eds.), *Second International Handbook of Mathematics Education* (pp. 323-349). Dordrecht, the Netherlands: Kluwer Academic Publishers.
Huygens, C. (1895). *Oeuvres complètes de Christiaan Huygens* (Vol. VI). Den Haag, the Netherlands: Nijhoff.
Iamblichus (1991). Iamblichus' mathematical work. In *Greek Mathematics* (Vol. 1). Cambridge, MA: Harvard University Press.
Jacobs, V. R. (1999). How do students think about statistical sampling before instruction? *Mathematics Teaching in the Middle School, 5*, 240-263.
Janvier, C. (Ed.). (1987). *Problems of representation in teaching and learning mathematics*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Jones, G. A., Langrall, C. W., Thornton, C. A., Mooney, E. S., Wares, A., Jones, M. R., et al. (2001). Using students' statistical thinking to inform instruction. *Journal of Mathematical Behavior, 20*, 109-144.
Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B., & Putt, I. J. (2000).

A framework for characterizing children's statistical thinking. *Mathematical Thinking and Learning, 2*, 269-307.
Kadunz, G. (1998). Visualisierung, Bild und Metapher. Die vermittelnde Tätigkeit der Visualisierung beim Lernen von Mathematik [Visualizing, image, and metaphors. The mediating activity of visualizing when learning mathematics]. *Journal für die Mathematik der Didaktik*, 280-302.
Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237-251.
Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition, 11*, 131-141.
Kanselaar, G., Van Galen, F., Beemer, H., Erkens, G., & Gravemeijer, K. P. E. (1999). *Grafieken leren met de computer [Learning graphs with the computer]*. Utrecht: Universiteit Utrecht/Onderwijskunde/ISOR.
Kant, I. (1787/1971). *Kritik der reinen Vernunft [Critique of pure reason]*. Hamburg, Germany: Meiner.
Keijzer, R. (2003). *Teaching formal mathematics in primary education; fraction learning as mathematising process*. Utrecht, the Netherlands: CD Beta Press.
Kelly, I. W., & Beamer, J. E. (1986). Central tendency and dispersion: The essential union. *Mathematics Teacher, 79*(1), 59-65.
Kelly, A. E., & Lesh, R. (Eds.). (2000). *Handbook of research design in mathematics and science education*. Mahwah, NJ: Lawrence Erlbaum Associates.
Kendall, M. G. (1960). Where shall the history of statistics begin? *Biometrika, 47*, 447-449.
Kendall, M. G. (1968). Certainty about uncertainty. *Statistica Neerlandica, 22*(1), 1-12.
Kiaer, A. N. (1895). Observations et expériences concernant des dénombrements réprésentatifs. *Bulletin of the International Statistical Institute, 9*(2), 176-183.
King, A. C., & Read, C. B. (1963). *Pathways to probability. History of mathematics of certainty and chance*. New York: Holt, Rinehart and Winston.
Kirshner, D., & Whitson, J. A. (Eds.). (1997). *Situated cognition theory: Social, semiotic, and neurological perspectives*. Hillsdale, NJ: Erlbaum.
Klaassen, C. W. J. M. (1995). *A problem-posing approach to teaching the topic of radioactivity*. Utrecht: CD Bèta Press.
Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6*, 59-98.
Konold, C. (1994). Understanding probability and statistical inference through resampling. In L. Brunelli & G. Cicchitelli (Eds.), Proceedings of the First Scientific Meeting of the International Association for Statistical Education (pp. 199-211). Perugia, Italy: Università di Perugia.
Konold, C., & Higgins, T. (2002). Highlights of related research. In S. J. Russell, D. Schifter & V. Bastable (Eds.), *Developing mathematical ideas: Working with data* (pp. 165-201). Parsippany, NJ: Dale Seymour Publications.
Konold, C., & Higgins, T. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 193-215). Reston, VA: National Council of Teachers of Mathematics.
Konold, C., & Khalil, K. (2003). *If u can graff these numbers – 2, 15, 6 – your stat literit.* Paper presented at AERA, Chicago.
Konold, C., & Miller, C. (2004). Tinkerplots. Data analysis software for middle school curricula. San Francisco: Key Curriculum Press.
Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*, 259-289.
Konold, C., Pollatsek, A., & Well, A. D. (1997). Students analyzing data: Research of critical barriers. In J. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 151-167). Voorburg, the Netherlands: International Statistical Institute.
Konold, C., Robinson, A., Khalil, K., Pollatsek, A., Well, A. D., Wing, R., et al. (2002). Stu-

dents' use of modal clumps to summarize data. In B. Phillips (Ed.), *Proceedings of the International Conference on Teaching Statistics [CD-ROM], Cape Town*. Voorburg, the Netherlands: International Statistics Institute.
Krüger, L., Daston, L. J., & Heidelberger, M. (Eds.). (1989). *The probabilistic revolution.* Ideas in history (Vol. 1). Cambridge, MA : MIT Press.
Lacan, J. (1968). *The language of the self; The function of language in psychoanalysis* (translated, with notes and commentary, by Anthony Wilden). Baltimore: Johns Hopkins University Press.
Lajoie, S. P., & Derry, S. J. (Eds.) (1993). *Computers as cognitive tools.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Lajoie, S. P. (Ed.). (1998). *Reflections on statistics: Learning, teaching, and assessment in grades K-12*. Mahwah, NJ: Lawrence Erlbaum Associates.
Laplace, P. S. (1812). *Théorie analytique des probabilités. Oevres complètes Vol. 7 (1891)*. Paris: Gauthier-Villar.
Latour, B. (1990). Drawing things together. In M. Lynch & S. Woolgar (Eds.), *Representation in scientific practice* (pp. 19-68). Cambridge, MA: MIT Press.
Lehrer, R., Schauble, L., Carpenter, S., & Penner, D. (2000). The interrelated development of inscriptions and conceptual understanding. In P. Cobb, E. Yackel & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms; Perspectivies on discourse, tools, and instructional design*. Mahwah, NJ: Lawrence Erlbaum Associates.
Lehrer, R., & Romberg, T. (1996). Exploring children's data modeling. *Cognition and Instruction, 14*, 69-108.
Lehrer, R., & Schauble, L. (2001). *Accounting for contingency in design experiments.* Paper presented at AERA, Seattle, WA.
Lemke, J. L. (2003). Mathematics in the middle: Measure, picture, gesture, sign, and word. In M. Anderson, A. Sáenz-Ludlow, S. Zellweger & V. V. Cifarelli (Eds.), *Educational perspectives on mathematics as semiosis: From thinking to interpreting to knowing* (pp. 215-234). Ottawa: Legas Publishing.
Lesh, R., & Doerr, H. M. (Eds.). (2003). *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching*. Mahwah, N.J: Lawrence Erlbaum Associates.
Lesh, R., & Kelly, A. E. (Eds.) (2000). *Research design in mathematics and science education*. Hillsdale, NJ: Erlbaum.
Letsios, D. G. (1996). *Nomos Rhodion Nautikos; Das Seegesetz der Rhodier. Untersuchungen zu Seerecht und Handelsschiffahrt in Byzanz [Rhodian sea law. Investigations into the sea law and trade sailing in Byzantium]* (Vol. 1). Rhodos: Aliki Kiantou - Pambouki.
Lewin, K. (1951). *Field theory in social sciences; Selected theoretical papers* (Cartwright, D., Ed.). New York: Harper and Row.
Lowndes, R., & Rudolf, G. R. (Eds.). (1975). *General average and York Antwerp rules* (Tenth ed. Vol. 7). London: Stevens & Sons.
Makar, K., & Confrey, J. (in press). Secondary teachers' statistical reasoning in comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecht, the Netherlands: Kluwer Academic Publishers.
Maso, I., & Smaling, A. (1998). *Kwalitatief onderzoek: praktijk en theorie [Qualitative research: Practice and theory]*. Amsterdam: Boom.
McClain, K. (2000). An analysis of the teacher's role in supporting the emergence of symbolizations in one first-grade classroom. *Journal of Mathematical Behavior, 19*, 189-207.
McClain, K. (2002). Teacher's and students' understanding: The role of tools and inscriptions in supporting effective communication. *Journal of the Learning Sciences, 11*, 217-249.
McClain, K., & Cobb, P. (2001). Supporting students' ability to reason about data. *Educational Studies in Mathematics, 45*, 103-129.
McClain, K., McGatha, M., & Hodge, L. L. (2000). Improving data analysis through discourse. *Mathematics Teaching in the Middle School, 5*, 548-553.

McGatha, M. (1999). *Instructional design in the context of developmental research: Documenting the learning of a research team*.Unpublished manuscript, Nashville, TN.

McGatha, M., Cobb, P., & McClain, K. (2002). An analysis of students' initial statistical understanding: Developing a conjectured learning trajectory. *Journal of Mathematical Behavior, 21*, 339-355.

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations in box plots. *The American Statistician, 32*(1), 12-16.

McNamara, O. (2003). Locating Saussure in Contemporary Mathematics Education Discourse. In M. Anderson, A. Sáenz-Ludlow, S. Zellweger & V. V. Cifarelli (Eds.), *Educational Perspectives on Mathematics as Semiosis: From Thinking to Interpreting to Knowing* (pp. 17-34). Ottawa: Legas Publishing.

Meira, L. (1995). Microevolution of mathematical representations in children's activity. *Cognition and Instruction, 13*, 269-313.

Meira, L. (1998). Making sense of instructional devices: The emergence of transparancy in mathematical activity. *Journal for Research in Mathematics Education, 29*, 121-142.

Meletiou, M. (2002). Conceptions of variation: A literature review. *Statistics Education Research Journal, 1*(1), 46-52. *(http://fehps.une.edu.au/serj)*

Meletiou, M., & Lee, C. (2002). Student understanding of histograms: A stumbling stone to the development of intuitions about variation. In B. Phillips (Ed.), *Developing a Statistically Literate Society. Proceedings of the Sixth International Conference on Teaching Statistics [CD-ROM]. Cape Town*. Voorburg, the Netherlands: International Statistics Institute.

Menne, J. (2001). *Met sprongen vooruit [Jumping ahead]*. Utrecht, the Netherlands: CD Beta Press.

Methodewijzer (1998). *Moderne wiskunde zevende editie methodewijzer*. Groningen, the Netherlands: Wolters Noordhoff.

Mevarech, Z. (1983). A deep structure model of students' statistical misconceptions. *Educational Studies in Mathematics, 14*, 415-429.

Mickelson, W. T., & Heaton, R. M. (in press). Primary teachers' statistical reasoning about data. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecht, the Netherlands: Kluwer Academic Publisher.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (Second ed.). Thousand Oaks, CA: Sage Publications.

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education, 26*, 20-39.

Monjardet, B. (1991). Élements pour une histoire de la médiane métrique [Elements of a history of the metric median]. In J. Feldman, G. Lagneau & B. Matalon (Eds.), *Moyenne, milieu, centre: histoire et usages* (pp. 45-62). Paris: Ecole des Hautes études en Sciences Sociales.

Moore, D. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of the giants: New approaches to numeracy* (pp. 95-137). Washington, DC: National Academy Press.

Moore, D. S. (1997). New pedagogy and new content: The case for statistics. *International Statistical Review, 65*, 123-165.

Mortensen, C., & Roberts, L. (1997). Semiotics and the foundations of mathematics. *Semiotica, 115*(1&2), 1-25.

Muller, F., & Thiel, J. H. (1986). *Beknopt Grieks-Nederlands woordenboek [Concise Greek-Dutch dictionary]* (11 ed.). Groningen, the Netherlands: Wolters-Noordhoff.

NCTM (1989). *Curriculum and evaluation standards*. Reston, VA: National Council of Teachers of Mathematics.

NCTM (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Neeleman, W., & Verhage, H. (1999). Hoogtepunten van grafische verwerking [Highlights of graphical representation]. *Nieuwe Wiskrant, 18*(3), 18-33.

Nelissen, J. M. C. (1987). *Kinderen leren wiskunde; Een studie over constructie en reflectie in het basisonderwijs*. Gorinchem, the Netherlands: De Ruiter.
Nelissen, J. M. C. (1999). Alweer die ijsbeer [Again that polar bear]. *Willem Bartjens, 18*(3), 36-37.
Nemirovsky, R., & Monk, S. (2000). "If you look at it the other way..."; An exploration into the nature of symbolizing. In P. Cobb, E. Yackel & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms; Perspectivies on discourse, tools, and instructional design* (pp. 177-221). Mahwah, NJ: Lawrence Erlbaum Associates.
Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society, 97*, 558-606.
Nisbett, R. E., Krantz, D., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90*, 339-363.
Noss, R. (2001). For a learnable mathematics in the digital culture. *Educational Studies in Mathematics, 48*, 21-46.
Noss, R., & Hoyles, C. (1996). *Windows on mathematical meaning*. Dordrecht, the Netherlands: Kluwer Academic Publishers.
Noss, R., Pozzi, S., & Hoyles, C. (1999). Touching epistomologies: Meanings of average and variation in nursing practice. *Educational Studies in Mathematics, 40*, 25-51.
OECD (1999). *Measuring student knowledge and skills: A new framework for assessment*. Paris: Organization for Economic Cooperation and Development.
Otte, M. (1998). Limits of contructivism: Kant, Piaget and Peirce. *Science & Education, 7*, 425–450.
Otte, M. (2003). Complementary, sets and numbers. *Educational Studies in Mathematics, 53*, 203-228.
Pannekoek, A. (1961). *A history of astronomy*. London: Allen and Unwin.
Pea, R. D. (1987). Cognitive technologies for mathematics education. In A. Schoenfeld (Ed.), *Cognitive science and mathematics education* (pp. 89-122). Hillsdale, NJ: Lawrence Erlbaum Associates.
Peirce, C. S. (1894). What is a sign? In P. E. Project (Ed.), *The essential Peirce; Selected philosophical writings* (Vol. 2 (1893-1913), pp. 4-10). Bloomington & Indianapolis, IN: Indiana University Press.
Peirce, C. S. (NEM). *The new elements of mathematics* (Eisele, C., Ed.) (Vol. I-IV). The Hague-Paris/Atlantic Highlands, N.J.,: Mouton/Humanities Press.
Peirce, C. S. (CP). *Collected papers of Charles Sanders Peirce*. Cambridge, MA: Harvard University Press.
Peirce, C. S. (EP). *The essential Peirce; Selected philosophical writings* (Vol. 1 & 2). Bloomington & Indianapolis, IN: Indiana University Press.
Perry, M., & Kader, G. (1998). Counting penguins. *Mathematics Teacher, 91*, 110-116.
Petrosino, A. J., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning, 5*, 131-156.
Pfannkuch, M., & Wild, C. (in press). Towards an understanding of statistical thinking. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecht, the Netherlands: Kluwer Academic Publisher.
Pijls, M., Dekker, R., & Van Hout-Wolters, B. (2003). Mathematical level raising through collaborative investigations with the computer. *International Journal of Computers for Mathematical Learning, 8*, 191-213.
Pine, K. J., & Messer, D. J. (2000). The effect of explaining another's action; On children's implicit theories of balance. *Cognition and Instruction, 18*, 35-51.
Plackett, R. L. (1970). The principle of the arithmetic mean. In E. Pearson & M. G. Kendall (Eds.), *Studies in the History of Statistics and Probability* (Vol. 1). London: Griffin.
Plackett, R. L. (1988). Data analysis before 1750. *International Statistical Review, 56*, 181-195.

Plön, O., & Kreutziger, G. (1965). *Das Recht der grossen Haverei [The law of general average]* (Vol. 1). Hamburg, Germany: Otto Meissner Verlag.
Pollatsek, A., Lima, S., & Well, A. D. (1981). Concept or computation: Students' understanding of the mean. *Educational Studies in Mathematics, 12*, 191-204.
Porter, T. M. (1986). *The rise of statistical thinking, 1820-1900*. Princeton, NJ: Princeton University Press.
Portney, S., & Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statistical Science, 12*, 279-300.
Presmeg, N. (1998). Metaphoric and metonymic signification in mathematics. *Journal of Mathematical Behavior, 17*, 25-32.
Presmeg, N. (2002). Transitions in emergent modeling. In K. P. E. Gravemeijer, R. Lehrer, B. van Oers & L. Verschaffel (Eds.), *Symbolizing, modeling and tool use in mathematics education* (pp. 131-138). Dordrecht, the Netherlands: Kluwer Academic Publishers.
Proust, M. (1923/1929). The captive (Vol. 5 of Remembrance of things past; C. K. Scott Moncrieff, Trans.). Retrieved from http://gutenberg.net.au/ebooks03/0300501.txt on February 23, 2004.
Ptolemy (1998). *Almagest* (G. J. Toomer, Trans.). Princeton, NJ: Princeton University Press.
Pyzdek, T. (2001). *The Six Sigma handbook*. New York: McGraw-Hill.
Rabinovitch, N. L. (1973). *Probability and statistical inference in ancient and medieval Jewish literature*. Toronto: University of Toronto Press.
Radford, L. (2000). Historical formation and student understanding of mathematics. In J. Fauvel & J. van Maanen (Eds.), *History in mathematics education: The ICMI study*. Dordrecht, the Netherlands: Kluwer Academic Publishers.
Reading, C., & Shaughnessy, J. M. (2000). Student perceptions of variation in a sampling situation. In T. Nakahara & H. Koyama (Eds.), *Proceedings of the 24th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 89-96). Hiroshima, Japan: Department of Mathematics Education, Hiroshima University.
RAC (Research Advisory Committee) (1996). Justification and reform. *Journal for Research in Mathematics Education, 27*, 516-520.
Roth, W.-M. (1996). Where is the context in contextual word problems? Mathematical practices and products in grade 8 students' answers to story problems. *Cognition and Instruction, 14*, 487-527.
Roth, W.-M. (2003). Competent workplace mathematics: How signs become transparent in use. *International Journal of Computers for Mathematical Learning, 8*, 161-189.
Roth, W.-M., & Bowen, G. M. (2001). Professionals read graphs: A semiotic analysis. *Journal for Research in Mathematics Education, 32*, 159-194.
Roth, W.-M., & McGinn, M. K. (1998). Inscriptions: Toward a theory of representing as social practice. *Review of Educational Research, 68*(1), 35-59.
Rubin, A., Bruce, B., & Tenney, Y. (1990). *Learning about sampling: Trouble at the core of statistics.* Paper presented at the Third International Conference on Teaching Statistics, Dunedin, New Zealand.
Rubin, E. (1968). The statistical world of Herodotus. *The American Statistician, 1968* (February), 31-33.
Rubin, E. (1971). Quantitative commentary on Thucydides. *The American Statistician, 1971* (December), 52-54.
Russell, S. J., & Corwin, R. B. (1989). *Statistics: The shape of the data. Used numbers: Real data in the classroom. Grades 4-6*. Washington, DC: National Science Foundation.
Sáenz-Ludlow, A. (2003). Classroom discourse in mathematics as an evolving interpreting game. In M. Anderson, A. Sáenz-Ludlow, S. Zellweger & V. V. Cifarelli (Eds.), *Educational perspectives on mathematics as semiosis: From thinking to interpreting to knowing* (pp. 253-281). Ottawa, Canada: Legas Publishing.
Salomon, G. (1998). Novel constructivist learning environments and novel technologies: Some issues to be concerned with. *Research Dialogue in Learning and Instruction, 1*(1),

3-12.
Saussure, F. de (1916/1974). *Course in general linguistics*. Fontana: Collins.
Schifter, D., & Fosnot, C. T. (1993). *Reconstructing mathematics education*. New York: Teachers College Press.
Schwartz, D. L., Goldman, S. R., Vye, N. J., Barron, B. J., & Cognition and Technology Group at Vanderbilt (1998). Aligning everyday and mathematical reasoning: The case of sampling assumptions. In S. P. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K-12* (First ed., pp. 233-273). Mahwah, NJ: Lawrence Erlbaum Associates.
Schwartzman, S. (1994). *The words of mathematics: An etymological dictionary of mathematical terms used in English*. Washington, DC: Mathematical Association of America.
Seeger, F. (1998). Representations in the mathematics classroom: Reflections and constructions. In F. Seeger, J. Voigt & U. Waschescio (Eds.), *The culture of the mathematics classroom* (pp. 308-343). Cambridge: Cambridge University Press.
Seeger, F. (2001). *Learning as acquisition and construction - Potentials of a global semiotic perspective*. Presented at EARLI.
Seeger, F. (2002). *Peirce and Vygotskij: A semiotic neighbourhood.* Paper presented at the Fifth Congress of the International Society for Cultural Research and Activity Theory. 18-22 June 2002, Amsterdam.
Sfard, A. (1991). On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics, 22*, 1-36.
Sfard, A. (2000a). Steering (dis)course between metaphors and rigor: Using focal analysis to investigate an emergence of mathematical objects. *Journal for Research in Mathematics Education, 31*, 296-327.
Sfard, A. (2000b). Symbolizing mathematical reality into being. In P. Cobb, E. Yackel & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms. Perspectives on discourse, tools, and instructional design* (pp. 37-98). Mahwah, NJ: Lawrence Erlbaum Associates.
Shafir, E., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition, 49*, 11-36.
Shaughnessy, J. M., Garfield, J., & Greer, B. (1996). Data handling. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 205-237). Dordrecht, the Netherlands: Kluwer.
Shaughnessy, J. M., Watson, J. M., Moritz, J., & Reading, C. (1999). *School mathematics students' acknowledgment of statistical variation.* NCTM Research Presession Symposium: There's more to life than centers. Paper presented at the 77th Annual NCTM Conference, San Francisco, California.
Sherin, M. G., Louis, D., & Mendez, E. (2000). Students' building on one another's mathematical ideas. *Mathematics Teaching in the Middle School, 6*, 186-190.
Sheynin, O. (1996). *The history of the theory of errors*. Egelsbach: Verlag Dr. Markus Hänsel-Hohenhausen.
Simon, J. L., & Bruce, P. (1991). Resampling: A tool for everyday statistical work. *Chance: New Directions for Statistics and Computing, 4*, 22-32.
Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26*, 114-145.
Simpson, J. A., & Weiner, E. S. C. (Eds.). (1989). *Average. In: The Oxford English dictionary* (Second ed. Vol. 1). Oxford, United Kingdom: Clarendon Press.
Singer, J. D., & Willett, J. B. (1990). Improving the teaching of applied statistics: Putting the data back into data analysis. *The American Statistician, 44*, 223-230.
Sittig, J., & Freudenthal, H. (1951). *De juiste maat: Lichaamsafmetingen van Nederlandse vrouwen als basis van een nieuw maatsysteem voor damesconfectiekleding [The right size: Body measures of Dutch women as a basis for a new size system of lady ready-to-wear clothes]*. Leiden: Stafleu.

Snee, R. D. (1990). Statistical thinking and its contribution to total quality. *The American Statistician, 44*, 116-121.

Spruit, J. E. e. a. (Ed.). (1996). *Corpus iuris civilis; Tekst en vertaling [Latin text and translation]* (Vol. III). Zutphen, the Netherlands: Walburg Pers.

Stamhuis, I. H. (1996). Christiaan Huygens correspondeert met zijn broer over levensduur [Christiaan Huygens corresponds with his brother about life span]. *De Zeventiende Eeuw, 12*(1), 161-170.

Stamhuis, I. H., & Koetsier, T. (1991). 'Die God tergen hebben onzekerheden'. *Wijsgerig Perspectief, 32*(6), 162-168.

Steffe, L. P., & Thompson, P. W. (2000). Teaching experiments methodology: Underlying principles and essential elements. In R. Lesh & A. E. Kelly (Eds.), *Research design in mathematics and science education* (pp. 267-307). Hillsdale, NJ: Erlbaum.

Steinbring, H. (1980). *Zur Entwicklung des Wahrscheinlichkeitsbegriffs - Das Anwendungsproblem in der Wahrscheinlichkeitstheorie aus didaktischer sicht [On the development of the probability concept –The applicability problem in probability theory from a didactical perspective]*. Bielefeld: Institut für Didaktik der Mathematik der Universität Bielefeld.

Steinbring, H. (1997). Epistemological investigation of classroom interaction in elementary mathematics teaching. *Educational Studies in Mathematics, 32*, 49-92.

Stigler, S. M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885-1920. *Journal of the American Statistical Association*.

Stigler, S. (1977). Eight centuries of sampling inspection: The trial of the Pyx. *Journal of the American Statistical Association, 72*, 439-500.

Stigler, S. M. (1980). R.H. Smith, a Victorian interested in robustness. *Biometrika, 67*, 217-221.

Stigler, S. M. (1984). Boscovich, Simpson and a 1760 manuscript on fitting a linear relation. *Biometrika, 71*, 615-620.

Stigler, S. M. (1986). *The history of statistics. The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Stigler, S. M. (1999). *Statistics on the table. The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.

Stjernfelt, F. (2000). Diagrams as centerpiece of a Peircean epistemology. *Transactions of the Charles S. Peirce Society, 36*, 357-384.

Strauss, A., & Corbin, J. (1988). *Basics of qualitative research. Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: SAGE Publications.

Strauss, S., & Bichler, E. (1988). The development of children's concepts of the arithmetic mean. *Journal for Research in Mathematics Education, 19*, 64-80.

Streefland, L. (1991). *Fractions in realistic mathematics education: A paradigm of developmental research*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Székely, G. (1997). Problem corner. *Chance, 10*(4), 25.

Tall, D., Thomas, M., Davis, G., Gray, E., & Simpson, A. (2000). What is the object of the encapsulation of a process? *Journal of Mathematical Behavior, 18*, 223-241.

Thucydides. (1954). *History of the Peloponnesian war* (Warner, R., Trans.). Baltimore: Penguin Books.

Thucydides. (1975). *Peri tou Peloponnesiakou polemou [On the Peloponnesian war]* (R. Schlatter, Ed.; T. Hobbes, Trans.). New Brunswick, NJ: Rutgers University Press.

Treffers, A. (1987). *Three dimensions. A model of goal and theory description in mathematics instruction - The Wiskobas project*. Dordrecht, the Netherlands: Reidel Publishing Company.

Treffers, A., Streefland, L., & De Moor, E. (1994). *Proeve van een nationaal programma voor het reken-wiskundeonderwijs op de basisschool. Deel 3A breuken [Attempt to a national program for mathematics education at primary school. Part 3a fractions]*. Tilburg, the Netherlands: Zwijsen.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tufte, E. R. (2000). *Visual explanations: Images and quantities, evidence and narrative*. Cheshire, CT: Graphics Press.
Tufte, E. R. (2001). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110.
Tversky, A., & Kahneman, D. (1982). Belief in the law of small numbers. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 23-31). New York: Cambridge University Press.
Tversky, A., & Shafir, E. (1992). The disjunction effect in choice under uncertainty. *Psychological Science, 3*, 305-309.
Tzou, C., & Cobb, P. (2000). *Supporting students' understanding of the relationships between data creation and data analysis.* Paper presented at the American Educational Research Association, New Orleans.
Van Amerom, B. A. (2002). *Reinvention of early algebra. Developmental research on the transition from arithmetic to algebra*. Utrecht, the Netherlands: CD Bèta Press.
Van Brummelen, G. (1998). Mathematical Methods in the Tables of Planetary Motion in Kushyar ibn Labban's Jami Zij. *Historia Mathematica, 25*, 265-280.
Van den Akker, J. (1999). Principles and methods of development research. In J. van den Akker, R. M. Branch, K. Gustafson, N. Nieveen & T. Plomp (Eds.), *Design approaches and tools in education and training* (pp. 1-14). Boston, Dordrecht: Kluwer Academic Publishers.
Van den Boer, C. (2003). *Als je begrijpt wat ik bedoel. Een zoektocht naar mogelijke verklaringen voor achterblijvende prestaties van allochtone leerlingen in het wiskundeonderwijs [If you know what I mean. A search for an explanation of lagging results of mathematics education among ethnic minority students]*. Utrecht, the Netherlands: CD Beta Press.
Van den Brink, F. J. (1989). *Realistisch Rekenonderwijs aan Jonge Kinderen [Realistic mathematics education to young children]*. Utrecht: OW&OC.
Van den Heuvel-Panhuizen, M. (1993). Toetsontwikkelingsonderzoek [Assessment design research]. In R. de Jong & M. Wijers (Eds.), *Ontwikkelingsonderzoek; theorie en praktijk [Design research; theory and practice]* (pp. 85-109). Utrecht, the Netherlands: Freudenthal Institute.
Van den Heuvel-Panhuizen, M. (1996). *Assessment and realistic mathematics education*. Utrecht, the Netherlands: CD Beta Press.
Van den Heuvel-Panhuizen, M. (Ed.). (2001). *Children learn mathematics; A learning-teaching trajectory with intermediate attainment targets*. Utrecht, the Netherlands: Freudenthal Institute.
Van Densen, W. L. T. (2001). *On the perception of time trends in resource outcome: Its importance in fisheries co-management agriculture and whaling*. Enschede, the Netherlands: Twente University.
Van der Hoeven, P. (1854). *Handleiding tot het opmaken van de avarijen [Guide to calculate averages]*. Dordrecht, the Netherlands: P.K. Braat.
Van Dijk, I. M. A. W., Van Oers, B., & Terwel, J. (2003). Providing or designing? Constructing models in primary maths education. *Learning and Instruction, 13*, 53-72.
Van Hiele, P. M. (1974). Het ontwerpen van een vertikale leerstofplanning voor de statistiek [The design of a vertical curriculum plan for statistics]. *Euclides, 49*, 247-251.
Van Oers, B. (2000). The appropriation of mathematics symbols: A psychosemiotic approach to mathematics learning. In P. Cobb, E. Yackel & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms; Perspectivies on discourse, tools, and instructional design* (pp. 133-176). Mahwah, NJ: Lawrence Erlbaum Associates.
Van Someren, M., Reimann, P., Bozhimen, H., & De Jong, T. (Eds.). (1998). *Learning with multiple representations*. Oxford, United Kingdom: Elsevier Science Ltd.
Véron, J., & Rohrbasser, J.-M. (2000). Lodewijk et Christiaan Huygens: La distinction entre

vie moyenne et vie probable [The distinction between mean and median life time]. *Mathématiques et Sciences Humaines, 38*(149), 7-21.

W12-16. (1992). *Achtergronden van het nieuwe leerplan wiskunde 12-16 [Backgrounds of the new curriculum of mathematics 12-16]*. Utrecht, the Netherlands: Freudenthal Institute.

Walker, H. M. (1931). *Studies in the history of statistical methods with special reference to certain educational problems*. Baltimore: Williams & Wilkins Company.

Walkerdine, V. (1988). *The mastery of reason*. London: Routledge.

Wassell, S. R. (2002). Rediscovering a family of means. *The Mathematical Intelligencer, 24*(2), 58-65.

Watson, J. M. (2002). Creating cognitive conflict in a controlled research setting: Sampling. In B. Phillips (Ed.), *Developing a Statistically Literate Society; Proceedings of the Sixth International Conference of Teaching Statistics Cape Town [CD-ROM]*. Voorburg, the Netherlands: International Statistics Institute.

Watson, J. M. (in press). Developing reasoning about samples. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Watson, J. M., & Kelly, B. A. (2002). Can grade 3 students learn about variation? In B. Phillips (Ed.), *Developing a statistically literate society. Proceedings of the sixth international conference on teaching statistics, Cape Town [CD-ROM]*. Voorburg, the Netherlands: International Statistics Institute.

Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*, 145-168.

Watson, J. M., & Moritz, J. B. (2000). Developing concepts of sampling. *Journal for Research in Mathematics Education, 31*, 44-70.

Wenger, R. H. (1998). *Communities of practice: Learning, meaning, and identity*. New York: Cambridge University Press.

Weytsen, Q. (1641). *Een tractaet van avarien, dat is gemaekt by Quintyn Weytsen.*

Whitson, J. A. (1997). Cognition as a semiotic process: From situated mediation to critical reflective transcendence. In D. Kirshner & J. A. Whitson (Eds.), *Situated cognition theory: Social, semiotic, and neurological perspectives* (pp. 97-150). Hillsdale, NJ: Erlbaum.

Whitson, J. A. (2003). *Thinking, learning, knowing, doing, and becoming: Semiotic and pragmatic bases for an alternative to "transfer"*. Paper presented at AERA, Chicago.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistics Review, 67*, 223-265.

Wilden, A. (1968). Lacan and the discourse of the other. In Lacan, J., *The language of the self* (pp. 159-311). Baltimore: Johns Hopkins University Press.

Wilensky, U. (1997). What is normal anyway? Therapy for epistemological anxiety. *Educational Studies in Mathematics, 33*, 171-202.

Wilkinson, L. (1999). Dot plots. *The American Statistician, 53*, 276-281.

Wittgenstein, L. (1984). *Philosophische untersuchungen [Philosophical investigations]*. Frankfurt am Main, Germany: Suhrkamp.

Yackel, E. (2000). Introduction: Perspectives on semiotics and instructional design. In P. Cobb, E. Yackel & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms; Perspectives on discourse, tools, and instructional design* (pp. 1-13). Mahwah, NJ: Lawrence Erlbaum Associates.

Yackel, E., & Cobb, P. (1996). Sociomathematical norms, Argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education, 27*, 458-477.

Zawojewski, J. S., & Shaughnessy, J. M. (1999). Data and chance. In E. A. Silver & P. A. Kenney (Eds.), *Results from the seventh mathematics assessment of the National Assessment of Educational Progress* (pp. 235-268). Reston, VA: National Council of Teachers of Mathematics.

Zawojewski, J. S., & Shaughnessy, J. M. (2000). Mean and median: Are they really so easy? *Mathematics Teaching in the Middle School, 5*, 436-440.

Samenvatting

1 Inleiding

In de huidige informatiemaatschappij heeft iedere burger enige statistische geletterdheid nodig. In veel beroepen en takken van wetenschap is zelfs een goede statistische kennis vereist. Het onderwijs slaagt er over het algemeen niet in om leerlingen een goede statistische basis te geven. Zo blijkt uit veel buitenlands onderzoek dat leerlingen de statistische grafieken en begrippen die ze geleerd hebben meestal niet goed kunnen gebruiken bij het analyseren van data. Er is daarom behoefte aan een empirisch onderbouwde instructietheorie die aangeeft hoe leerlingen een bepaald domein kunnen leren. Het doel van het hier beschreven onderzoek is om bij te dragen aan zo'n instructietheorie voor aanvankelijke statistiek. Daartoe is al een aanzet gedaan door Cobb, McClain, Gravemeijer en hun team aan de Vanderbilt University, Nashville, TN, USA. Het onderhavige onderzoek is hier een vervolg op.
De methode die hier gehanteerd wordt is ontwikkelingsonderzoek (*design research*). Dit houdt kort gezegd in dat er in een cyclisch proces een leertraject wordt ontworpen, in klassensituaties getoetst en waar nodig aangepast op grond van retrospectieve analyses. In het begin van een ontwikkelingsonderzoek heeft de voorlopige instructietheorie nog een hypothetisch karakter, maar door de cyclische doordenking en beproeving wordt de zich ontwikkelende instructietheorie steeds beter empirisch onderbouwd. In dit onderzoek staan twee onderwerpen centraal: symboliseren en het gebruik van computertools.
Symboliseren. In de statistiek spelen grafieken een centrale rol, maar zonder geschikte concepten kunnen leerlingen deze niet goed interpreteren, selecteren of produceren. Vanuit een semiotisch perspectief geformuleerd is het zaak dat grafieken een symboolfunctie krijgen: ze moeten in de ogen van leerlingen gaan staan voor statistische objecten, bijvoorbeeld voor de frequentieverdeling van een dataverzameling. Onder de noemer van symboliseren wordt dit proces geanalyseerd.
Computertools. Zonder computerprogramma's is het analyseren van data een uiterst bewerkelijk proces. Omdat de meeste statistische programma's niet geschikt zijn om statistiek mee te *leren*, wordt in dit onderzoek gebruikgemaakt van statistische minitools, die binnen het Nashville-onderzoek speciaal voor leerlingen van ongeveer twaalf jaar zijn ontworpen.

2 Achtergrond en onderzoeksvragen

Een van de belangrijkste uitgangspunten van het onderzoek is dat wiskunde leren een betekenisvolle activiteit moet zijn. De theorie van het realistisch wiskundeonderwijs (RME) biedt hiervoor richtlijnen en ontwerpheuristieken. In plaats van kant-en-klare wiskunde aan te bieden die vervolgens wordt toegepast, wordt gestreefd naar een proces van geleid heruitvinden.

We hebben het onderzoek voorbereid in verschillende stappen. Na een literatuurstudie zijn de bevindingen van het Nashville-onderzoek samengevat. Ook is een historische studie gedaan naar de ontwikkeling van statistische concepten om ideeën op te doen voor de leergang. Vervolgens zijn exploratieve interviews gehouden met 26 Nederlandse brugklasleerlingen om het beginniveau voor de leergang vast te kunnen stellen. Dit geheel van bevindingen leidde tot een zogenaamde didactische fenomenologie die weer de basis vormde voor de eerste leergang. We vatten eerst de literatuurstudie en het Nashville-onderzoek samen.
Statistische data-analyse leidt tot de beschrijving en voorspelling van eigenschappen van *groepen* gegevens, niet van individuele gevallen. Uit de literatuurstudie van onderzoek naar statistiekonderwijs blijkt dat leerlingen het moeilijk vinden om kenmerken van een dataverzameling als een geheel te beschrijven. In plaats daarvan letten ze vooral op individuele gegevens. Verder blijken ze veel moeite te hebben met het interpreteren en gebruiken van grafieken waarin individuele gegevens samengevoegd zijn, zoals in een histogram of een boxplot.
Het kernbegrip in de statistiek waarmee patronen in variabele fenomenen en groepseigenschappen van dataverzamelingen gemodelleerd worden, is *verdeling*. Het uitgangspunt van het Nashville-onderzoek was dat leerlingen met een notie van verdeling beter in staat zouden zijn om groepseigenschappen te beschrijven en te voorspellen. Voor het onderzoek in de basisvorming denken we daarbij in het bijzonder aan de verdeling van data over een variabele. Omdat een formele definitie van bijvoorbeeld de normale verdeling te moeilijk is voor brugklasleerlingen, is er net als in het Nashville-onderzoek voor gekozen om de *vorm* van verdelingen centraal te stellen in het ontwerp van de leergang.

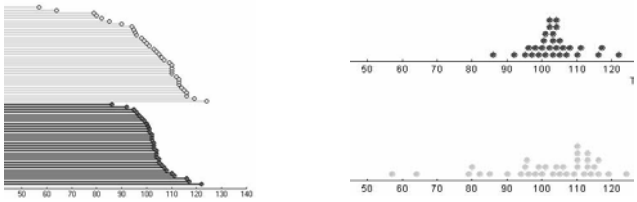


Figure S.1: Minitool 1 en 2 met dezelfde dataverzameling, de levensduur in uren van batterijen van twee merken

Voor het onderzoek in Nashville zijn drie Java-applets ontwikkeld die het leerproces van leerlingen moesten ondersteunen: de minitools. In de eerste minitool wordt iedere meetwaarde door een horizontale staaf weergegeven: de lengte is evenredig met de meetwaarde (Figuur S.1 links). In de tweede minitool zijn de eindpunten van die staven als het ware recht naar beneden gevallen. Hierdoor ontstaat een stippendiagram (Figuur S.1 rechts), waarin bijvoorbeeld een normaal verdeelde dataver-

zameling ook te zien is als de bekende klokvorm. In de derde minitool kunnen leerlingen een *scatterplot* op verschillende manieren structureren, maar deze tool is in het onderhavige onderzoek niet gebruikt.
In het Nashville-onderzoek is het inderdaad gelukt om leerlingen te laten redeneren over groepseigenschappen van verdelingen, bijvoorbeeld via de vorm ("de heuvel zit hier meer naar links", dus de meerderheid van de data was lager). Er waren echter ook nieuwe vragen, met name rond het gemiddelde en steekproeven, en het symboliseerproces. Op basis van de literatuurstudie en het Nashville-onderzoek is in het Nederlandse onderzoek gekozen voor de volgende onderzoeksvragen:

1 *Hoe kunnen leerlingen met weinig statistische achtergrond een begrip van verdeling ontwikkelen?*
2 *Hoe verloopt het proces van symboliseren als leerlingen over verdeling leren redeneren?*

3 Methodologie

Ontwikkelingsonderzoek bestaat in de regel uit cycli van drie fasen. In de eerste fase van een cyclus wordt een klassenexperiment voorbereid. Op basis van beschikbare kennis en ervaring wordt een hypothetisch leertraject uitgestippeld met een einddoel, lesmateriaal en verwachtingen over het leerproces van de leerlingen. In de tweede fase wordt het klassenexperiment uitgevoerd. Op basis van ervaringen in de klas kan het hypothetisch leertraject gedurende het klassenexperiment aangepast worden. In de derde fase wordt het leerproces geanalyseerd. Allerlei vragen kunnen onderzocht worden, bijvoorbeeld: zijn de verwachtingen uitgekomen? Welke aanpassingen zijn er nodig? Wat voor begrip van spreiding hebben leerlingen in een bepaalde context? Op basis van de retrospectieve analyse kan het hypothetisch leertraject gereviseerd worden voor een volgende cyclus. Het kenmerkende van ontwikkelingsonderzoek is dat steeds de meest kansrijke route gekozen wordt. Dit kan betekenen dat het hypothetisch leertraject aangepast wordt als het klassenexperiment nog niet is afgelopen. Er kunnen tijdens de tweede fase ook minicycli van analyse en revisie plaatsvinden.
Bij de retrospectieve analyse in klassen 1B (havo) en 2B (havo-vwo) hebben we bij episodes vermoedens geformuleerd die vervolgens aan de rest van de episodes en andere databronnen zijn getoetst. Dit proces van vermoedens genereren en toetsen is enkele keren herhaald volgens een methode die lijkt op de constant comparatieve methode van Glaser en Strauss (1967).
In het onderzoek zijn verschillende data verzameld om het leerproces van leerlingen te registreren: audio- en video-opnamen van leerlingen in de klas, leerlingmateriaal, toetsen en veldobservaties, maar ook audio-opnamen van reflectiegesprekken met de assistenten na afloop van iedere les (zie Tabel 3.3). Een belangrijk onderdeel van

het datacorpus vormde de verzameling mini-interviews die tot doel hadden om te achterhalen wat grafieken en begrippen voor leerlingen betekenden. De interviews werden gehouden in de klas aan de hand van vooraf vastgestelde vragen die met de assistenten werden doorgesproken. De lengte van de interviews varieerde van ongeveer twintig seconden tot vier minuten.
De audio- en video-opnamen van klassen 1B en 2B zijn getranscribeerd. De protocollen van 1B zijn in een computerprogramma voor protocolanalyse (Erkens, 2001) ingevoerd en gecodeerd op opgave, woorden en begrippen. Het doel hiervan was om de protocollen systematisch te analyseren, hypothesen te genereren en alle episodes die over een bepaald probleem of begrip gingen makkelijk terug te vinden om zo de hypothesen te testen. De protocollen van klas 2B zijn gecodeerd door de onderzoeker en deels met drie assistenten doorgesproken. Er was een hoge mate van overeenstemming over de codering.

4 Historische fenomenologie

Als voorbereiding op het ontwerp van een hypothetisch leertraject is onder andere een historische studie gedaan naar de ontwikkeling van enkele statistische representaties en kernbegrippen zoals gemiddelde, mediaan, steekproef en verdeling. Er is onderzocht welke fenomenen aanleiding gaven tot het ontstaan van statistische begrippen en hoe deze begrippen gebruikt zijn om fenomenen te organiseren. Op basis van de historische voorbeelden zijn hypothesen geformuleerd over het leren van die begrippen.
De eerste hypothese luidde dat schatten van grote aantallen een mogelijk geschikte context was om het gebruik van een impliciete notie van gemiddelde in relatie tot een totaal aantal te stimuleren. Een andere hypothese was dat leerlingen het bereikmidden, het gemiddelde van de twee extreme waarden, als centrummaat zouden gebruiken. Ook zijn er hypothesen over de mediaan, steekproef, verdeling en grafieken geformuleerd. De hypothesen die getoetst konden worden, werden bij de experimenten alle op één na bevestigd.
De historische analyse hielp het ontwerpproces op verschillende manieren. Ze bleek nuttig om aspecten van begrippen als het gemiddelde en steekproef te onderscheiden naar moeilijkheidsgraad, en hielp om beter door de ogen van leerlingen te kijken.

5 Exploratieve interviews en een didactische fenomenologie

Uit de analyse van de exploratieve interviews met 26 brugklassers bleek dat deze leerlingen het aritmetische gemiddelde redelijk goed konden berekenen. Ook verbonden ze aan het informele begrip ‘gemiddelde’ allerlei kwalitatieve eigenschappen zoals de meeste, ongeveer, balanspunt, bereikmidden, de middelste, zwaartepunt en meerderheid. Er leek echter een kloof te zijn tussen het aritmetische gemiddelde als algoritme en de kwalitatieve eigenschappen van het gemiddelde zoals dat in het dagelijks leven ook wel gebruikt wordt. Verder bleken twee voor-

beeldopgaven van het Nashville-onderzoek met minitool 1 en 2 ook voor de Nederlandse leerlingen geschikt te zijn.
Op basis van de literatuurstudie, de historische studie en exploratieve interviews is een analyse gemaakt van het begrip verdeling en andere statistische kernbegrippen. In lijn met het Nashville-onderzoek, werden leerlingen gestimuleerd om te leren beschrijven hoe data verdeeld zijn en uit de vorm van grafieken groepseigenschappen af te leiden. De verwachting was: als het begrip 'verdeling' voor leerlingen een objectkarakter krijgt, dan kunnen ze er ook eigenschappen van onderzoeken (centrum, spreiding, dichtheid, scheefheid). De vraag is alleen hoe het objectvormingsproces verloopt. De bestaande theorieën hierover gaan uit van een procedure die tot object wordt samengevat, maar in het geval van het begrip verdeling lijkt eerder sprake te zijn van een samengestelde eenheid (*composite unit*). Net als het getal 10 gedacht kan worden als tien eenheden en als een eenheid, kan een dataverzameling gezien worden als een collectie getallen maar ook als een eenheid met eigenschappen die de elementen van de collectie niet hebben.
Geïnspireerd door de historische voorbeelden van het gemiddelde dat vermoedelijk gebruikt werd om grote aantallen te schatten, besloten we de leergang met schatten te beginnen. We vroegen leerlingen eerst het aantal olifanten op een foto te schatten. Een verwante schattingsvraag was: hoeveel brugklasleerlingen mogen er in een ballon als er normaal gesproken acht volwassenen in mogen? Aan de hand van dergelijke vragen verwachtten we noties van gemiddelde en steekproef te kunnen bespreken.
Het hypothetisch leertraject, in één zin samengevat, was dat leerlingen zouden leren redeneren over verdelingsaspecten met steeds geavanceerdere grafieken en begrippen. De middelen die dit proces moesten ondersteunen waren de minitools, de activiteiten die in Nashville ontwikkeld waren, maar ook nieuw ontwikkelde onderwijsactiviteiten.

6 Een hypothetisch leertraject ontwikkelen voor de brugklas

In verschillende cycli is een hypothetisch leertraject ontwikkeld voor het laatste brugklasexperiment (in 1B). Binnen bepaalde contexten en in relatie tot staafgrafieken en stippendiagrammen leerden de leerlingen om allerlei aspecten van dataverzamelingen te beschrijven en te gebruiken in hun redeneringen: het gemiddelde, betrouwbaarheid (van batterijen), uitschieters, meer hoge waarden (scheve verdeling), verspreid of dicht bij elkaar. Om leerlingen te stimuleren zich op groepseigenschappen te richten, vroegen we hun om grafieken verzinnen die voldeden aan bepaalde groepseigenschappen, bijvoorbeeld een onbetrouwbaar batterijmerk met een hoge levensduur.
Een van de conclusies van de retrospectieve analyse was dat leerlingen geneigd zijn om verdelingen in te delen in lage, 'gemiddelde' en hoge waarden (hypothese C1 in

de Appendix). Verder bleek het nuttig om leerlingen zelf grafieken te laten maken en te laten vergelijken, en ze 'wat-als-vragen' te stellen over hypothetische situaties. In klas 1E (een vwo-klas) ontstond tijdens een klassengesprek over twee leerlinggrafieken een discussie over de bult die in een van de grafieken te zien was (Figuur S.3 links). Dit was de eerste keer dat er over de vorm van een verdeling geredeneerd werd. Leerlingen uit die klas gebruikten het begrip 'bult' vervolgens ook om andere problemen op te lossen. Omdat deze klas het einddoel van redeneren over verdeling middels haar vorm het dichtst genaderd is, hebben we het symboliseerproces in deze klas in een apart hoofdstuk (8) geanalyseerd. Verder besloten we om het einddoel 'verdeling als object met eigenschappen' op te geven voor 1B, omdat we ervan overtuigd waren geraakt dat dit einddoel te ambitieus was voor een havo-klas in slechts twaalf lessen. In plaats daarvan leek het verstandiger om genoegen te nemen met het leren beschrijven hoe data verdeeld zijn en om de begrippen spreiding en steekproef centraler te stellen.
Door verschillende cycli van ontwikkelingsonderzoek te doorlopen is een hypothetisch leertraject ontwikkeld dat gefundeerd is op patronen in het leerproces van leerlingen en de leermiddelen die dit ondersteunen. De verschillen tussen de verschillende klassenexperimenten laten zien hoe subtiele wijzigingen in de vraagstelling kunnen leiden tot aanzienlijke verschillen in de leeropbrengst.

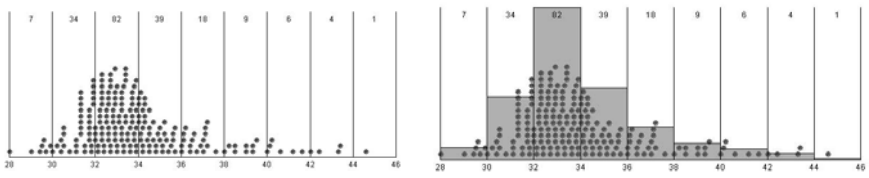


Figure S.2: Het spijkerbroekenprobleem: taillegegevens van 200 mannen in inches. Hier is gekozen voor vaste intervalbreedte (links), waarna het in de nieuwe versie van minitool 2 mogelijk is om voor een histogram te kiezen (rechts).

7 Het hypothetisch leertraject toetsen in een brugklas

Het leerproces in 1B bleek voor het grootste deel te verlopen zoals was voorspeld in het hypothetisch leertraject dat in de voorgaande experimenten ontwikkeld was. Een uitzondering hierop waren de activiteiten rond de mediaan, omdat de discussie over het nut van de mediaan te abstract bleek te zijn. Een andere uitzondering was het spijkerbroekenprobleem dat diende als voorbereiding op het histogram als symbool voor een frequentieverdeling (Figuur S.2): met name de rol van steekproeven in het spijkerbroekenprobleem bleek moeilijk te zijn. Het bescheidener einddoel om te kunnen beschrijven hoe data verspreid zijn, pakte wel goed uit. Leerlingen hadden een lokale visie op spreiding die mogelijk voorbereidde op een dichtheidsnotie en

een verdelingsbegrip: bij het uitleggen hoe data verspreid waren, beschreven leerlingen vaak de verdeling van de data.
Tijdens het laatste brugklasexperiment ontstond een nieuwe onderzoeksactiviteit: groeiende steekproeven. Bij het voorspellen van grafieken van grotere steekproeven spraken leerlingen van een preciezer gemiddelde, van een grotere spreidingsbreedte, meer 'gemiddelde' waarden, en voorspelden ze redelijk vloeiende en adequate vormen voor grote steekproeven. De activiteit rond groeiende steekproeven lokte coherent redeneren over statistische noties inclusief vorm uit. Daarom is besloten dit idee in een tweede klas te gebruiken als een steeds terugkerend thema.
Met de resultaten van de brugklasexperimenten kon de eerste onderzoeksvraag beantwoord worden. In het kort: brugklasleerlingen kunnen leren redeneren over verdeling op een informele manier als een vergelijkbaar hypothetisch leertraject gebruikt wordt. Desondanks raden we niet aan om verdeling als einddoel te formuleren in de basisvorming: het begrip is lastig om te ontwikkelen en te toetsen. Bovendien zijn de gunstige omstandigheden van het onderzoek (inclusief de mini-interviews) niet representatief voor reguliere onderwijssituaties. Het lijkt wel haalbaar voor leerlingen in de basisvorming om een statistische taal ontwikkelen waarin ze kunnen uitdrukken hoe data verspreid zijn.

8 Diagrammatisch redeneren met de 'bult'

In klas 1E is het redeneren over de vorm van verdelingen het dichtst bij het einddoel gekomen. Dit redeneren is semiotisch geanalyseerd om inzicht te krijgen in het proces van symboliseren en zo de tweede onderzoeksvraag te kunnen beantwoorden. Er zijn verschillende theorieën over betekenisgeving toegepast op de episodes waarin met bulten werd geredeneerd. Het vruchtbaarst als analyse-instrument bleek de semiotiek van Peirce te zijn. Het belangrijkste voordeel van Peirces semiotiek boven bijvoorbeeld ketens van betekenisgeving (afkomstig van Lacan) was het niet-lineaire aspect. Met name de begrippen diagrammatisch redeneren en hypostatische abstractie bleken nuttig om het symboliseerproces te analyseren.
Diagrammatisch redeneren bestaat uit drie stappen: een diagram maken, ermee experimenteren en reflecteren op de resultaten. Bij het reflecteren is het van belang dat eigenschappen van diagrammen beschreven worden (*predikatie*). Een eigenschap van een diagram (bijvoorbeeld: "de stipjes zijn erg verspreid") kan vervolgens als een zelfstandig object gezien worden dat weer eigenschappen heeft ("de spreiding is groot"). Deze abstractiestap, waarbij een predikaat tot een nieuw object wordt gemaakt, noemt Peirce *hypostatische abstractie*. Zo werd de beschrijving van "veel rond het gemiddelde en wat minder kleine en grote waarden" uiteindelijk benoemd met 'bult': een object dat beschreven kon worden en waarmee geredeneerd werd.

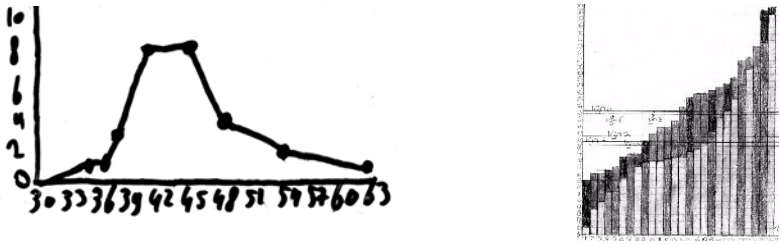


Figure S.3: De leerlinggrafieken die aanleiding gaven om over bulten te redeneren (links Mikes en rechts Emily's grafiek)

De 'bult' was niet alleen maar een plaatje of een metafoor: al snel konden leerlingen uitleggen waardoor de bult in Mikes grafiek veroorzaakt werd en waarom die bult in Emily's grafiek er als een horizontaal recht stuk uitzag. In dit stadium representeerde de bult voor leerlingen de grote groep waarden in het midden (de 'meerderheid'). In de volgende les redeneerden leerlingen met het begrip 'bult' om twee verdelingen te vergelijken: omdat de bult van één merk hoger lag, vonden leerlingen dat merk beter. De bult functioneerde dus als een groepsrepresentant, net zoals het gemiddelde vaak gebruikt wordt in de statistiek. In de volgende les zijn vragen gesteld om leerlingen te stimuleren met de bult als een object te laten opereren: wat gebeurt er met de bult als we niet brugklassers wegen maar tweedeklassers? Leerlingen verschoven de bult inderdaad als geheel ("dan is de bult meer naar rechts") en meenden dat die dezelfde vorm bleef houden als de steekproef steeds groter zou worden. Leerlingen hebben dus de gewichtdata in een bult gesymboliseerd: de bult is symbool gaan staan voor de verdeling van de data. Hier is het symboliseerproces dus geslaagd.
Aan de hand van de drie stappen van diagrammatisch redeneren konden we nog meer conclusies trekken. In de eerste stap, het diagrammatiseren, is het belangrijk dat leerlingen zelf diagrammen maken die voor hen betekenisvol zijn, of representaties zoals de minitools gebruiken die ze makkelijk kunnen interpreteren. Software is vooral nuttig bij de experimenteerfase van diagrammatisch redeneren: het experimenteren met redelijk omvangrijke dataverzamelingen is ondoenlijk met de hand, maar eenvoudig met de software. Verder kunnen leerlingen de data op verschillende manieren organiseren met de minitoolopties, zoals sorteren op grootte en subgroep, en verschillende groepen maken. De derde stap, de reflectie, is uiterst belangrijk om predikatie en hypostatische abstractie te stimuleren. Opvallend was dat de beste redeneringen plaatsvonden tijdens klassendiscussies buiten het computerlokaal. De docent speelt een uiterst belangrijke rol bij het ensceneren van dergelijke discussies en bij het vormen van normen, bijvoorbeeld dat data gebruikt moeten worden als ze beschikbaar zijn. Het vergelijken van diagrammen van dezelfde dataverzameling kan zeer vruchtbaar zijn, vooral als leerlingen gestimuleerd worden te beschrijven

wat het gemeenschappelijke object is waarvan de diagrammen representaties zijn. Dit object is over het algemeen een hypostatische abstractie, letterlijk een nieuw object dat verondersteld wordt ten grondslag te liggen aan verschillende verschijningsvormen.

9 Diagrammatisch redeneren over groeiende steekproeven

In het vervolgexperiment in een tweede klas waren groeiende steekproeven een steeds terugkerend thema. Het doel was om het vermoeden te toetsen dat het redeneren over groeiende steekproeven het redeneren over de vorm van verdelingen in samenhang met steekproeven en andere statistische concepten zou stimuleren. In het tweede-klasexperiment is dit vermoeden bevestigd. Het hypothetisch leertraject luidde samengevat: progressief diagrammatisch redeneren over verdelingsaspecten in relatie tot groeiende steekproeven. Door deze formulering was het mogelijk om de twee onderzoeksvragen tegelijk te beantwoorden als antwoord op de volgende vraag:

Hoe kunnen leerlingen met weinig statistische kennis een begrip van verdeling ontwikkelen door diagrammatisch redeneren over groeiende steekproeven?

In de vierde les beschreven leerlingen allerlei eigenschappen van hun voorspellingsgrafieken in vergelijking tot grafieken met echte data (predikatie). Daarbij gebruikten ze allerlei termen die gaandeweg meer een objectkarakter kregen. Prototypisch is de overgang van “de stipjes zijn verspreid” naar “de spreiding is groot”. Bij de voorspelling van de gewichtsgrafiek van alle tweedeklassers in Utrecht dachten leerlingen aan drie vormen: piramide (Figuur S.4, links), klokvorm (rechts) en halve cirkel.

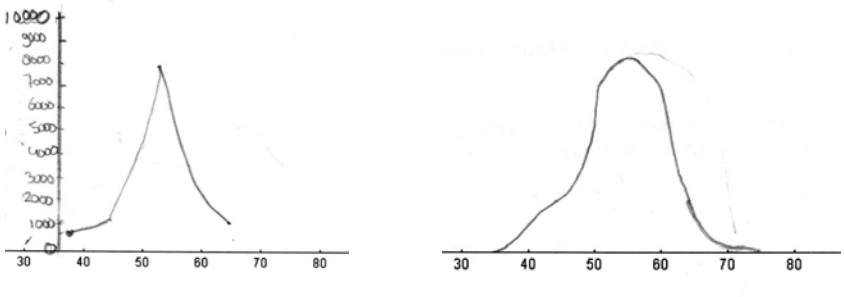


Figure S.4: Twee voorspellingen van de gewichtsverdeling van tweedeklassers

In de zesde les hebben we deze drie vormen plus twee scheve verdelingen als uitgangspunt genomen voor een discussie over de vraag welke het best bij de verdeling van gewicht past. De leerlingen participeerden goed in de discussie en gebruik-

ten allerlei statistische noties als instrumenten in hun redeneringen.
In de achtste les bleek dat veel leerlingen de steun van een bekende context nog niet konden missen. Ze zagen bijvoorbeeld lang niet altijd de vormen in groeiende steekproeven die ze uit tabellen met getallen trokken (de verdelingen waren uniform, normaal of scheef). Hoewel die vormen soms ook niet goed zichtbaar waren, speelde er vermoedelijk ook iets anders mee: de hypostatische abstractie van de verdelingsvormen was bij de leerlingen nog niet zo gevorderd was dat ze door de ruis van de variatie het signaal van de vormen zagen. Bovendien kwam aan het licht dat we te weinig aandacht hadden besteed aan het schalen van de *x*-as; dit was een missend element in de experimenteerfase met de minitools.
In de slotinterviews slaagden leerlingen er redelijk in om gemiddelde, mediaan en modus te lokaliseren in de schets van een scheve verdeling. Op de vraag wat een verdeling is, antwoordden enkelen: “hoe het verspreid is over de grafiek”. In overeenstemming met een conclusie van het experiment in 1B, bleek dat het begrip ‘spreiding’ voor leerlingen dicht tegen dat van verdeling aan lag (C7 in de Appendix). Anderen spraken over hoeveel stippen bepaalde waarden op de as kregen: “deze waarde krijgt wat meer en die waarde krijgt bijna niets”.
Dit geeft een mogelijk antwoord op de vraag uit hoofdstuk 5 wat de procedurele kant van het begrip verdeling zou kunnen zijn: data toevoegen aan de grafiek oftewel een steekproef laten groeien. Uit de analyses blijkt dat er vele stappen van hypostatische abstractie zijn in de ontwikkeling van het begrip verdeling. De vastgestelde gelaagdheid in het begrijpen van verdelingen kan helpen om bij een volgend experiment het einddoel ‘verdeling als object’ nauwkeuriger te operationaliseren.

10 Conclusies en discussie

Het doel van het onderzoek was om bij te dragen aan een empirisch onderbouwde instructietheorie voor aanvankelijk statistiekonderwijs. Er zijn hypothetische leertrajecten ontwikkeld voor de brugklas en de tweede klas, en er zijn patronen in het leerproces van leerlingen gevonden die in de meeste experimenten vergelijkbaar waren. De gemeenschappelijke patronen in het leerproces en de middelen die dit proces ondersteunen vormen de basis voor de instructietheorie. De antwoorden op de hoofdvragen kunnen als volgt worden samengevat.
1) Leerlingen kunnen onder gunstige omstandigheden een informeel begrip van verdeling ontwikkelen als de hypothetische leertrajecten uit dit onderzoek aangepast worden aan lokale omstandigheden. Met name de activiteiten rond groeiende steekproeven blijken het redeneren over verdeling te stimuleren. We raden echter niet aan om het redeneren over verdeling als vorm of object in de basisvorming als eindterm te kiezen.
2) Het proces van symboliseren kan beschreven worden als progressief diagrammatisch redeneren. Het is van belang dat leerlingen zelf diagrammen maken, dat ze ex-

perimenteren met data en diagrammen, bijvoorbeeld met geschikte computertools, en reflecteren op de resultaten. In de reflectiefase is het wenselijk dat het lesmateriaal en de docent gelegenheden creëren voor predikatie en hypostatische abstractie.
Het onderzoek heeft ook tot allerlei aanbevelingen geleid over het leren van statistische kernbegrippen, over symboliseren en het gebruik van computertools. We noemen enkele.

Data en context. Data zijn getallen in context (Moore, 1990). Als leerlingen zich niet goed in de context verdiepen, zijn ze geneigd met de data te rekenen, maar als ze zich alleen in de context verdiepen en de beschikbare data niet gebruiken, doen ze uitspraken die niet op de data gebaseerd zijn. Gezamenlijk doorspreken waar data vandaan komen is belangrijk om de connectie tussen data en context te leggen. Het is daarom onder meer wenselijk grote opdrachten te gebruiken waarbij leerlingen de context goed leren kennen. Nu krijgen leerlingen in reguliere wiskundelessen soms wel acht verschillende contexten per les voorgeschoteld, wat oppervlakkigheid in de hand kan werken (cf. Van den Boer, 2003).

Centrum van een verdeling. Het is van belang dat leerlingen een taal ontwikkelen waarmee ze hun intuïties en observaties in relatie tot simpele grafieken kunnen verwoorden. We kunnen gebruikmaken van de intuïtie die leerlingen al lijken te hebben van verdelingen: er zijn normale en uitzonderlijke gevallen, dus lage, gemiddelde en hoge waarden. De 'gemiddelde' groep kunnen we zien als een intuïtieve voorloper van het centrum van de verdeling. Pas als leerlingen met een notie van centrum kunnen redeneren lijkt het zin te hebben om dat centrum te meten met een formele maat zoals het gemiddelde of de mediaan.

Van spreiding naar verdeling. Als voorloper van het beschrijven van verdelingen, is het zinvol om leerlingen te laten verwoorden hoe data verspreid zijn. De begrippen 'bereik' en 'spreidingsbreedte' kunnen in een vroeg stadium aangeleerd worden om te voorkomen dat het woord 'spreiding' degenereert tot 'spreidingsbreedte'. Het kost meer moeite om tot een formele spreidingsmaat te komen. De meest voor de hand liggende spreidingsmaat is de kwartielafstand, maar die moet goed voorbereid worden, bijvoorbeeld via vier (ongeveer) even grote groepen of met een optie voor de middelste 50 procent van de data. Er lijkt echter een conflict te zijn tussen de intuïtie van een verdeling bestaande uit drie groepen en de formele manier om kwartielen te gebruiken, die tot vier groepen leiden. In ieder geval moet de rol van de mediaan opnieuw overdacht worden; de historische analyse kan daarvoor aanwijzingen bieden.

Groeiende steekproeven. Leerlingen waren vaak geneigd om erg kleine steekproeven te kiezen of de hele populatie te willen meten. De activiteiten rond groeiende steekproeven blijken een geschikte manier te zijn om leerlingen over steekproefgrootte en de vorm van een verdeling te laten redeneren. Bovendien kan middels dergelijke activiteiten het verband tussen steekproef en populatie gelegd worden.

Diagrammatiseren en symboliseren. Om diagrammatiseren en daarmee symboli-

seren te stimuleren is het zinvol om leerlingen zelf grafieken te laten verzinnen en die te vergelijken. Bij het voorspellen van hypothetische situaties zijn leerlingen gedwongen om vanuit een eigenschap van een dataverzameling als geheel te denken, zodat ze niet aan individuele data kunnen denken. Het is ook belangrijk om simpele grafieken aan te bieden die instrumenten in het diagrammatisch redeneren kunnen worden; staafgrafieken en bolletjesgrafieken (*dot plots*) zijn geschikte kandidaten. Sommige eigenschappen van histogrammen en boxplots zijn weliswaar makkelijk af te lezen, maar zulke geaggregeerde grafieken zijn voor leerlingen in de basisvorming lastig om te interpreteren als symbolen voor verdelingen. Als onderdeel van het diagrammatisch redeneren is het ook van groot belang dat de objecten van aandacht goed gedefinieerde onderwerpen van gesprek worden. Het is verder wenselijk dat leerlingen een wat-als-houding ontwikkelen en op de merites van verschillende datarepresentaties reflecteren. Het moeten aannemen van de rol van data-analist kan een dergelijke houding stimuleren.

Computertools. De minitools zijn simpele computertools die vrijwel geen technische problemen opwerpen, maar hier staat tegenover dat ze beperkte mogelijkheden bieden. Minitool 1, bijvoorbeeld, biedt alleen horizontale staven, terwijl veel leerlingen uit zichzelf verticale tekenen. De computertools lijken vooral nuttig te zijn om met dataverzamelingen en representaties te experimenteren. Het is belangrijk dat reflectie gestimuleerd wordt, en dit blijkt makkelijker te gaan als er geen computers binnen handbereik zijn. De ervaring leert dat de docent een uiterst belangrijke rol speelt bij het ondersteunen van de meeste van onze aanbevelingen. In algemenere zin is het onze overtuiging dat alle onderwijsfactoren zoals docentgedrag, leerstof, onderwijsactiviteiten en einddoelen op het gebruik van computertools afgestemd moeten worden. Alleen dan lijkt het de moeite waard om erin te investeren.

Curriculum vitae

Arthur Bakker was born in Hilversum, the Netherlands, on January 3, 1970. He completed his secondary education at the Rijksscholengemeenschap Noord-Kennemerland in Alkmaar. In 1995, the University of Amsterdam awarded him a Master's Degree in Mathematics (cum laude) with Philosophy and German as Minors. In Switzerland, Arthur investigated the foundations of set theory, in particular the work of Paul Finsler. The next academic year, 1996-1997, was devoted to the teacher training course at the University of Amsterdam and in August 1997, Arthur became a mathematics teacher at the Spinoza Lyceum, a Dalton school in Amsterdam.
Since November 1998, he has been working at the Freudenthal Institute, first as a Ph.D. student *(onderzoeker in opleiding)*, and later as team member of several projects: Tinkerplots (an NSF-granted project led by Cliff Konold), Mathematics in Context (revision of American mathematics textbooks), and Special Education.
From September 2004 to August 2006, he will be working as a Research Officer at the Institute of Education, University of London, with Celia Hoyles, Richard Noss, and Phillip Kent in a research project about techno-mathematical literacies in the workplace.