Statistics Education Research Journal

Volume 7 Number 2 November 2008

International Association for Statistical Education
http://www.stat.auckland.ac.nz/~iase

International Statistical Institute
http://isi.cbs.nl/

STATISTICS EDUCATION RESEARCH JOURNAL
The *Statistics Education Research Journal* (*SERJ*) is a peer-reviewed electronic journal of the International Association for Statistical Education (IASE) and the International Statistical Institute (ISI). *SERJ* is published twice a year and is free.

SERJ aims to advance research-based knowledge that can help to improve the teaching, learning, and understanding of statistics or probability at all educational levels and in both formal (classroom-based) and informal (out-of-classroom) contexts. Such research may examine, for example, cognitive, motivational, attitudinal, curricular, teaching-related, technology-related, organizational, or societal factors and processes that are related to the development and understanding of stochastic knowledge. In addition, research may focus on how people use or apply statistical and probabilistic information and ideas, broadly viewed.

The *Journal* encourages the submission of quality papers related to the above goals, such as reports of original research (both quantitative and qualitative), integrative and critical reviews of research literature, analyses of research-based theoretical and methodological models, and other types of papers described in full in the Guidelines for Authors. All papers are reviewed internally by an Associate Editor or Editor, and are blind-reviewed by at least two external referees. Contributions in English are recommended. Contributions in French and Spanish will also be considered. A submitted paper must not have been published before or be under consideration for publication elsewhere.

Further information and guidelines for authors are available at: http://www.stat.auckland.ac.nz/serj

Submissions

Manuscripts must be submitted by email, as an attached Word document, to co-editor Tom Short <tshort@jcu.edu>. Submitted manuscripts should be produced using the Template file and in accordance with details in the Guidelines for Authors on the Journal's Web page: http://www.stat.auckland.ac.nz/serj

Publication: IASE/ISI, Voorburg, The Netherlands
Technical Production: California Polytechnic State University, San Luis Obispo, California, United States of America
Web hosting and technical support: Department of Statistics, University of Auckland, New Zealand

ISSN: 1570-1824

SERJ EDITORIAL BOARD

TABLE OF CONTENTS

EDITORIAL

I am very happy to introduce this special issue of *SERJ* focusing on Informal Inferential Reasoning (sometimes referred to by the acronym IIR). This is my first editorial as co-editor of the *Journal* and when I sat down to write this editorial I found that most of my work had already been done for me by our two very capable guest editors, Dave Pratt and Janet Ainley. Their introduction explains the meaning of the term, gives the genesis of the special issue, and summarises the content of the papers: I don't feel that I need write any more about these aspects.

It has been a pleasure working with Dave and Janet, and I wish to thank them for the huge amount of work they have carried out in putting this issue of *SERJ* together. I would also like to acknowledge the work of a large group of people who have supported this process. First, of course, thanks to all the authors of the papers for sharing the results of their research and experience. Thanks also to the referees, who have reviewed the papers and given their suggestions for improvements, and to Carol Blumberg, Roxy Peck, and Chris Reading, who provided extra copy-editing assistance. And finally, a huge debt of gratitude to my previous co-editor, Iddo Gal, who first set this special issue in motion, and who passed it over to me at the beginning of this year in such a form that the work since then has been remarkably trouble free.

Having worked as co-editor for almost a year, I realise that I have joined a very capable team, including my co-editor Tom Short who has handled the various paper submissions and their refereeing, Assistant Editor Beth Chance who has put together both of the issues this year and overseen their proofing, and our team of Associate Editors, too numerous to mention by name. However, particular thanks to Dick Sheaffer, who is stepping down, and a welcome to Randall Groth, who is joining the editorial board as Associate Editor.

Now it's time to start thinking about the next special issue of our journal! We have had a few suggestions for the next theme, but would welcome any others while we are at this initial stage. But in the meantime, please enjoy the results of all our labours on this issue!

PETER PETOCZ

INTRODUCING THE SPECIAL ISSUE ON INFORMAL INFERENTIAL REASONING

GUEST EDITORS:

DAVE PRATT
Institute of Education, University of London, UK
dave.pratt@ioe.ac.uk

JANET AINLEY
University of Leicester, UK
janet.ainley@leicester.ac.uk

Inference is a foundational area in statistics, and learning and teaching about inference is a key concern of statistics education. The aim of this special issue is to advance the current state of research-based knowledge about the development, learning, and teaching of a critical subset of issues in this broad area; we focus on ***informal*** aspects of inferential reasoning. This topic was the focus of the Fifth International Forum on Statistical Reasoning, Thinking and Literacy, held at the University of Warwick, UK, in August 2007, and a number of the papers in this Special Issue have been developed from papers presented at that conference.

In selecting papers for the special issue, we have aimed to focus on learners' informal ideas about statistical inference or on people's intuitive ways of reasoning about statistical inference in diverse contexts rather than on mastery of formal procedures or methods. The first paper, by Allan Rossman, introduces the topic by setting out a statistician's view of the roots of statistical inference, emphasizing how those roots might be addressed by informal methods at the tertiary level. Although such approaches have relevance to secondary teaching, the intention is to provide a basis for the meaningful development of formal ideas. Subsequent papers in this special issue consider epistemological, psychological, and pedagogical dimensions that underpin informal inferential reasoning at all phases of education and beyond.

We recognize at the outset that the definition of what counts as "informal inference" is slippery: What is informal could depend on the nature of the inferential tasks being studied, on the complexity of the statistical or probabilistic concepts involved, on the educational stage, and on other factors. We see the two papers which follow Rossman's as framing the remainder of the special issue. Beyth-Marom, Fidler, and Cumming propose the term, *statistical cognition*, to unify the enterprise of mounting evidence across what have traditionally been the disparate disciplines of statisticians, psychologists, and educators. Zeiffler, Garfield, delMas, and Reading seek to build a working definition of informal inferential reasoning as *the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples.* Starting from this definition, the authors offer a framework for designing tasks for its study.

The remaining papers do indeed provide data that we might view as the starting point of such research and the reader might wish to consider these papers in the light of the framework proposed by Zeiffler et al. One might ask whether it is possible to locate these four studies in that framework or indeed whether the papers throw light upon the validity

Statistics Education Research Journal, 7(2), 3-4, http://www.stat.auckland.ac.nz/serj

or scope of the framework itself. At the same time, it is appropriate to ask whether such evidence supports the notion of *statistical cognition* as proposed by Beyth-Marom et al.

For example, Watson examines the informal inferential reasoning of 12- to 13-year-olds using *Tinkerplots*™. One claim in the paper is that the ease of creating representations in *TinkerPlots* may contribute to the students' developing intuitions about what might be considered as a *real* difference when comparing data sets. Paparistodemou and Meletiou-Mavrotheris also report on a classroom-based study involving *Tinkerplots*, though in their case at the younger age of 8 years. In support of the claim by Watson, they argue that the dynamic statistical software, in the context of carefully-focused questioning by the teacher, facilitated informal inferential reasoning. In both these papers the interactions between the design of tasks, the teacher's role, the functionality of the software, and children's ownership of the data are explored and discussed. In terms of the Beyth-Marom analysis, both provide evidence in relation to statistical cognition, and, with respect to Zeiffler et al., both respond to their call for "more research to explore the role of foundational concepts, data sets and problem contexts, and technology tools in helping students to reason informally, and then formally, about statistical inference."

The paper by Pratt, Johnston-Wilder, Ainley, and Mason extends the discussion to the context of probability and the responses of 10- to 11-year-olds to dice-throwing tasks. By drawing on Mason's theory of the Structure of Attention, the analysis provides a micro-level analysis of children's inferential reasoning. Whereas Watson and Paparistodemou and Meletiou-Mavrotheris explore students' inferential reasoning about survey data, Pratt et al. consider informal inferential reasoning when the data are generated by a die. As statisticians, we might think of the die's generational capacity as an instantiation of a theoretical uniform distribution. In comparing these three papers, we might ask how students' inferential reasoning is shaped by such a fundamental difference in the statistical structure of the task, the nature of the distribution as a collection of data, or as a theoretical potential.

In the final paper, Bakker, Kent, Derry, Noss, and Hoyles remind us that informal inferential reasoning is not the sole prerogative of education in school and college. They consider statistical process control in automotive manufacturing. The paper helps us to recognize the breadth of the notion of informal inferential reasoning, not only in terms of the age and context of the learners but also in terms of its epistemological location. Bakker et al. highlight the need to foreground the goals within the settings that structure the reasoning activity.

The collection of papers in this special issue illustrates then the importance of the topic. Research in this field, although still in its infancy, begins to offer insights into learners' inferential reasoning and how that thinking might be shaped more effectively by well-designed tasks. We see how inferential activity takes place at all ages and is important in itself and as a meaningful basis for a more formal treatment in later schooling. The issue as a whole points to the differing positions that future research on this topic might take: a developmental position in which learning can be assessed against an emerging framework; a micro-level position that focuses on students' inferential reasoning as it changes in the moment; an inferentialist position in which goals are related to the space of reasons within the setting.

REASONING ABOUT INFORMAL STATISTICAL INFERENCE: ONE STATISTICIAN'S VIEW

ALLAN J. ROSSMAN
California Polytechnic State University – San Luis Obispo
arossman@calpoly.edu

ABSTRACT

This paper identifies key concepts and issues associated with the reasoning of informal statistical inference. I focus on key ideas of inference that I think all students should learn, including at secondary level as well as tertiary. I argue that a fundamental component of inference is to go beyond the data at hand, and I propose that statistical inference requires basing the inference on a probability model. I present several examples using randomization tests for connecting the randomness used in collecting data to the inference to be drawn. I also mention some related points from psychology and indicate some points of contention among statisticians, which I hope will clarify rather than obscure issues.

Keywords: *Statistical reasoning; Statistical significance; Randomization tests*

1. PRELIMINARY DEFINITIONS

In preparing these comments I began by consulting an online dictionary (dictionary.com), where I found two definitions of "infer" that seem especially relevant to statistical inference:

1. to derive by reasoning; conclude or judge by premises or evidence;
2. to draw a conclusion, as by reasoning.

Similarly, the following two definitions of "informal" struck me as appropriate for this discussion:

1. without formality or ceremony, casual;
2. not according to the prescribed, official, or customary way or manner; irregular; unofficial.

I also consulted a statistics textbook, *The Statistical Sleuth* (Ramsey & Schafer, 2002), in which I read the following definitions:

1. An *inference* is a conclusion that patterns in the data are present in some broader context.
2. A *statistical inference* is an inference justified by a probability model linking the data to a broader context.

Informed by these definitions, I suggest that inference requires going beyond the data at hand, either by generalizing the observed results to a larger group (i.e., population) or by drawing a more profound conclusion about the relationship between the variables (e.g., that the explanatory variable causes a change in the response).

Statistical inference has traditionally been the focus of introductory courses at the tertiary level, and this topic has become more prevalent in the K-12 curriculum. For example, the K-12 GAISE (Guidelines for Assessment and Instruction on Statistics Education) report endorsed by the American Statistical Association (Franklin et al., 2005)

Statistics Education Research Journal, 7(2), 5-19, http://www.stat.auckland.ac.nz/serj

argues that by the end of their secondary schooling, students should learn to "look beyond the data." This GAISE report also emphasizes that these students should understand the nature of "chance variability."

This notion of *chance* variability is fundamental to drawing statistical inferences. Statisticians deliberately introduce randomness into the process of collecting data, in large part to enable inferences to be made on a probabilistic justification. This randomness takes one of two forms (or both), depending on the research question being addressed:

1. Random *sampling* from a population enables results about the sample to be generalized to the larger population.
2. Random *assignment* of units to treatment groups allows for cause-and-effect conclusions to be drawn about the relationship of the explanatory and response variables.

Figure 1, taken from *The Statistical Sleuth,* summarizes these points.

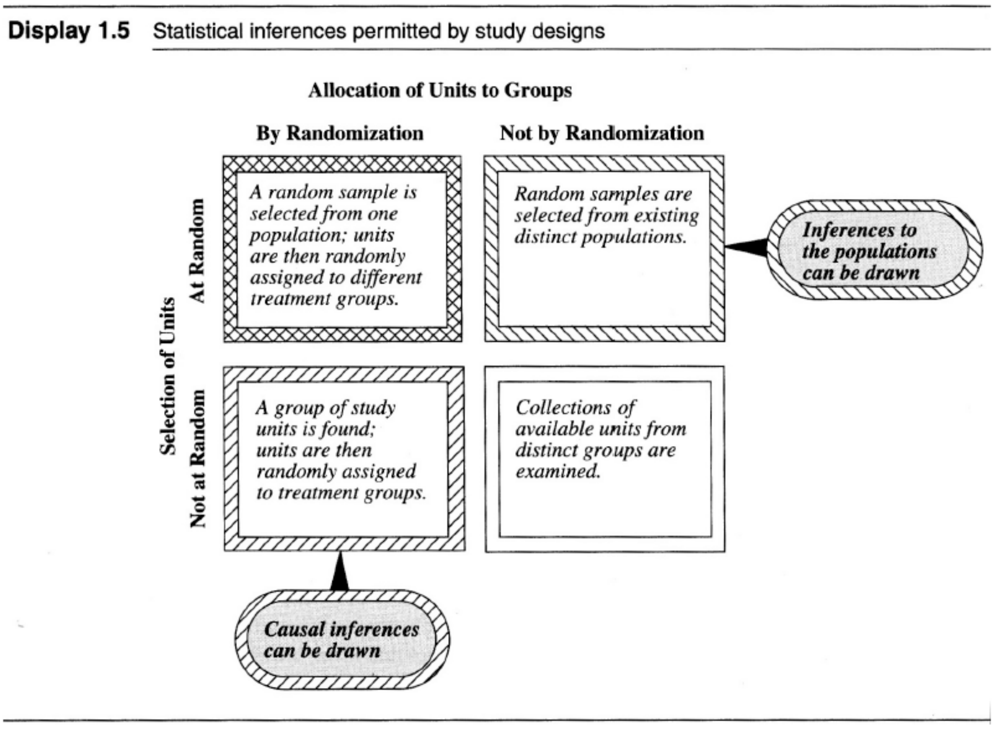

Display 1.5 Statistical inferences permitted by study designs
Allocation of Units to Groups
By Randomization
Not by Randomization
Selection of Units
At Random
Not at Random
A random sample is selected from one population; units are then randomly assigned to different treatment groups.
Random samples are selected from existing distinct populations.
Inferences to the populations can be drawn
A group of study units is found; units are then randomly assigned to treatment groups.
Collections of available units from distinct groups are examined.
Causal inferences can be drawn

Figure 1. Statistical inferences permitted by study designs from Ramsey and Shafer (2002)

2. AN EXAMPLE OF INTUITIVE INFERENTIAL REASONING

Example 0: Funny Dice Beth Chance, inspired by Jeff Witmer, introduced me to the following activity for introducing students to the reasoning of statistical significance. Take to class a pair of dice that appear to be fair and ordinary but are actually not: One of the dice contains only fives on its faces and the other has half twos and half sixes. The dice will therefore produce sums of only seven and eleven. (An internet search for "7 11 dice" will reveal many places to purchase such dice.) Roll the dice, or better yet ask a student to roll them, and call out the sum. Do this repeatedly, and observe the reactions of students in the class as the sevens and elevens accumulate.

Students generally have no reactions to the first two or three rolls. By the time they see a seven or eleven for the fourth or fifth time, some start to snicker or otherwise indicate that the results seem suspicious. By the sixth or seventh roll, many in the class openly voice conviction that the dice are not fair. After the tenth roll, almost all students in the class are convinced (without looking at the faces of the dice) that the dice are not fair. Most students provide a good account of their reasoning process, explaining that it would be extremely unlikely to observe so many results of seven or eleven if the dice were truly fair.

This reasoning process, which seems to come very naturally for students, is a classic example of Fisherian inductive reasoning. Students are assessing the strength of evidence against a claim. They do this by determining how unlikely the observed result would be, if in fact the claim being tested were true. Of course, all of this happens not just informally but intuitively. If I impose more structure here, I assert that students' intuitive reasoning process in this example involves:

- Starting with an unspoken belief that the dice are fair (we could call this the null model or null hypothesis);
- Evaluating that the observed data (nothing but sevens and elevens) would have been very unlikely if that belief (null model) were true (intuitively calculating a p-value);
- Rejecting the initial belief (null model) based on the very small p-value, rather than believe that a very rare event has occurred by chance alone.

3. AN EXAMPLE OF INFORMAL INFERENTIAL REASONING

Example 1: Toy Preference A recent study (Hamlin, Wynn, & Bloom, 2007) investigated whether infants take into account an individual's actions towards others in evaluating that individual as appealing or aversive, perhaps laying the foundation for social interaction. In one component of the study, 10-month-old infants were shown a "climber" character (a piece of wood with "google" eyes glued onto it) that could not make it up a hill in two tries. Then they were shown two scenarios for the climber's next try, one where the climber was pushed to the top of the hill by another character ("helper") and one where the climber was pushed back down the hill by another character ("hinderer"). Each infant was alternately shown these two scenarios several times. Then the child was presented with both pieces of wood (the helper and the hinderer) and asked to pick one to play with. Preferences were recorded for a sample of 16 infants, with 14 choosing the helper toy.

Clearly more than half of these infants chose the helper toy, but the inferential question is whether this result provides evidence of a *genuine* preference, either among a larger population of infants or among these same 16 infants if they were to be tested repeatedly. When asked for their initial impressions, students give widely varying reactions. Some are willing to conclude a genuine preference merely because more than half chose the helper; others argue that they would remain unconvinced about a genuine preference even if all 16 chose the helper toy because of what they perceive as a prohibitively small sample size.

Making a statistical inference requires a probability model. Fortunately, a simple model presents itself, one that is both familiar and understandable at the school level. Under the null model that infants have no genuine preference, we can model their selections as flips of a fair coin. In this manner we can simulate the selections by 16 infants over and over again, in order to assess how surprising it would be to obtain 14 or more of them choosing the helper in a sample of 16 infants if there were, in fact, no genuine preference. Asking students in a class to conduct 16 coin flips and count the number of heads, we quickly find that it is quite unusual to obtain 14 or more heads. Repeating this process 1000 times produces results like those in Figure 2.

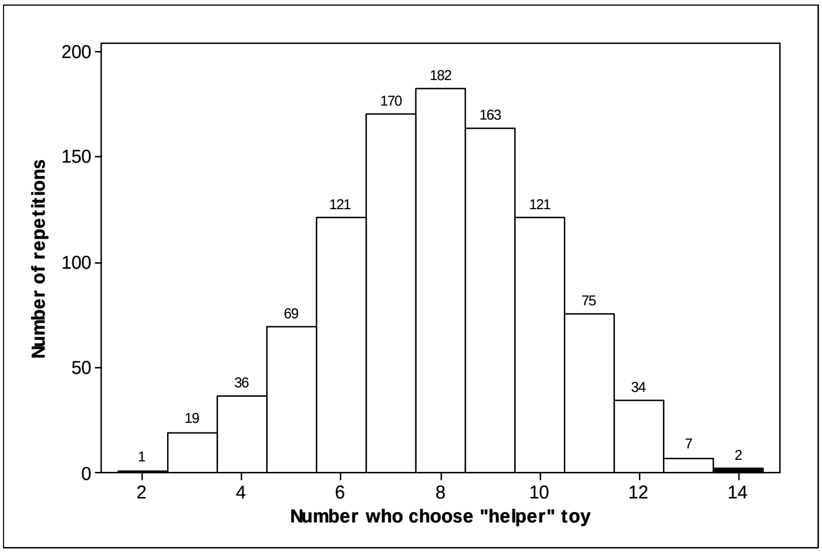


Figure 2. Simulation results for helper toy study

Notice that 14 chose the helper toy in 2 of these 1000 simulated repetitions. This graph therefore reveals that it is not impossible to find 14 or more choosing the helper toy even when there is no genuine preference, but such a result is very unlikely. Accordingly, we have strong evidence that infants genuinely do tend to prefer the helper toy over the hinderer.

This reasoning process is identical to that with the 7/11 dice, but it does not come as naturally to students. With the dice, intuition correctly tells us that it is very unlikely to get a string of exclusively sevens and elevens with fair dice. But our intuition is much less reliable for knowing the distribution of coin flip results. Simulation enables us to estimate the probability distribution and the p-value from this study. This activity and analysis are amenable to use with schoolchildren as well as college students.

But is this inferential analysis informal? I contend that it is. We are not doing a formal calculation of an exact p-value, which we could do using the binomial distribution. We are also not calculating a test statistic or approximate p-value based on a normal distribution. But we are using a reasonable process, based on a probability model, to draw an inference beyond the data at hand.

What other reasoning would I like students to think about, and begin to learn about in this context? Three issues, in order of increasing conceptual difficulty:

1. Students should recognize the key role that the sample proportion of successes plays in this inferential reasoning process. For example, if only 10 of the 16 infants had chosen the helper toy, this would provide much weaker support for concluding that infants have a genuine preference for the helper. Why? Because a result as extreme as 10 or more successes would not be at all surprising under the null model of no genuine preference (as the above histogram shows).
2. Students should come to appreciate the important role played by sample size. Ask about a different (hypothetical) study in which 100% choose the helper toy: Does that provide strong evidence of a genuine preference? Well, not if the study only

involved two infants. What if 60% choose the helper- is that statistically significant? Not if the study involved 10 infants, but yes if the study involved 100 infants.

3. Students should eventually learn to use the same reasoning process to investigate claims beyond a 50/50 null model. For example, do the study results provide evidence that infants actually prefer the helper by more than a 2-to-1 ratio over the hinderer? The reasoning process is the same, but the simulation needs to use a 2/3 probability of each infant's choosing the helper.

4. INFORMAL INFERENCE WITH RANDOMIZATION TESTS

Example 2: Dolphin Therapy Swimming with dolphins can certainly be fun, but is it also therapeutic for patients suffering from clinical depression? To investigate this possibility, researchers recruited 30 subjects aged 18-65 with a clinical diagnosis of mild to moderate depression (Antonioli & Reveley, 2005). Subjects were required to discontinue use of any antidepressant drugs or psychotherapy four weeks prior to the experiment, and throughout the experiment. These 30 subjects went to an island off the coast of Honduras, where they were randomly assigned to one of two treatment groups. Both groups engaged in the same amount of swimming and snorkeling each day, but one group did so in the presence of bottlenose dolphins and the other group (outdoor nature program) did not. At the end of two weeks, each subjects' level of depression was evaluated, as it had been at the beginning of the study. The results are summarized in Table 1 and Figure 3.

Table 1. Results of dolphin therapy experiment

	Dolphin therapy	Control group	Total
Showed substantial improvement	10	3	13
Did not show substantial improvement	5	12	17
Total	15	15	30

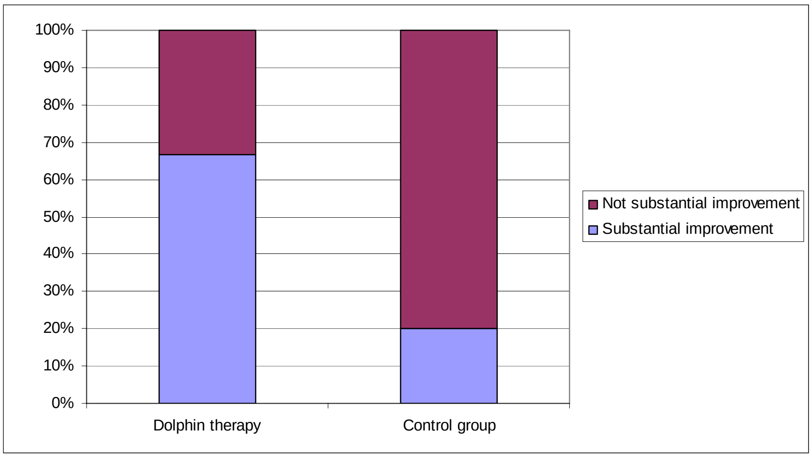


Figure 3. Results of dolphin therapy experiment

Clearly the dolphin therapy group had a larger success rate than the control group (66.7% vs. 20.0%). Can we reasonably *infer* that the dolphin therapy really is more effective than the control? To address this key question we must consider the role of chance variability.

The randomness in this study arises from researchers randomly assigning the 30 subjects to one of the treatment groups. Is it possible that this randomization process alone, even if dolphin therapy were no more effective than the control, would have produced results as extreme as the researchers found? Sure, it's possible. But is that possibility so unlikely that it discredits that explanation?

We could proceed directly to a probability calculation to assess this unlikeliness. But with introductory students I recommend investigating this with simulation. Students can simulate this randomization process by taking 30 playing cards, marking 13 to represent those who showed substantial improvement and the other 17 to represent those who did not improve substantially. Then shuffle the cards and randomly deal out 15 to be in the dolphin therapy group with the other 15 in the control group. Note that this shuffling/dealing process simulates the random assignment process actually used by the researchers to put subjects in treatment groups. Also note that this simulation process assumes that there is really no benefit of the dolphin therapy, because it assumes that the 13 subjects who improved were going to improve regardless of which group they were assigned to. Then observe the results of the simulated random assignment, either by calculating the difference in success proportions between the two groups or simply by noting the number of "successes" in the dolphin therapy group. Figure 4 shows the results of 1000 simulated random assignments.

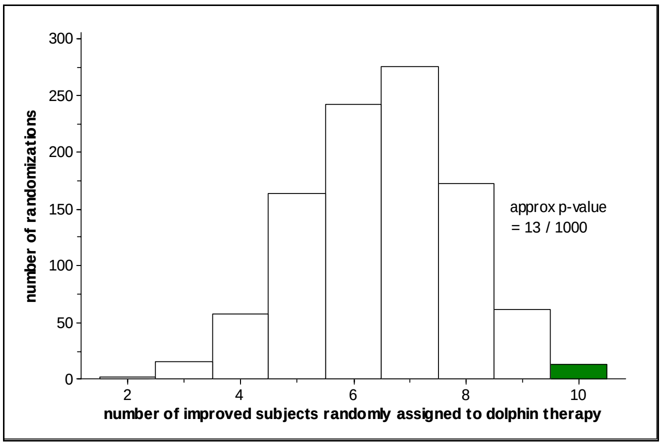


Figure 4. Simulation results from dolphin therapy experiment

In only 13 of the 1000 random assignments did the simulated result turn out to be as extreme as the actual experimental result (10 or more successes in the dolphin therapy group). So, it is indeed possible to have obtained such an extreme result by chance alone, even if the dolphin therapy had no effect, but this simulation reveals that this possibility is fairly unlikely. We therefore conclude that the experimental data provide fairly strong evidence that dolphin therapy really is more effective than the control.

The exact probability can be calculated to be 0.0127, to four decimal places. This procedure is known as Fisher's Exact Test. This is an example of a type of inference procedure called a randomization test. One advantage of this procedure for introducing introductory students to the reasoning process of statistical inference is that it makes clear the connection between the random assignment in the design of the study and the inference procedure. It also helps to emphasize the interpretation of a p-value as the long-term proportion of times that a result at least as extreme as in the actual data would have occurred by chance alone under the null model. For an overview of randomization tests, see Ernst (2004).

Notice that the reasoning process here is again the same as with the 7-11 dice, and also as with the helper/hinderer toy. But students seem to struggle more to understand it in this context. There are two possibilities (dolphin therapy is more effective than control, or it is not). The experimental data would be very unlikely to occur if dolphin therapy was not more effective, so we have strong evidence to reject that explanation and conclude that dolphin therapy really is more effective than the control. But of course this conclusion is not definitive, and there remains a possibility that dolphin therapy is not more effective and the researchers just happened to witness a rare event. Many students seem to be troubled by this lack of certainty in their conclusion, more so than with the 7-11 dice.

What else do I want students to learn about the reasoning of statistical inference in such settings? Similar to my list following Example 1, I want students to appreciate and understand the importance of how different the success rates are between the two groups, and also the importance of the numbers of subjects in the two groups.

Example 3: Murderous Nurse For several years in the 1990s, Kristen Gilbert worked as a nurse in the intensive care unit (ICU) of the Veteran's Administration hospital in Northampton, Massachusetts. Over the course of her time there, other nurses came to suspect that she was killing patients by injecting them with the heart stimulant epinephrine. Part of the evidence against Gilbert was a statistical analysis of more than one thousand 8-hour shifts during the time Gilbert worked in the ICU (Cobb & Gelbach, 2005). Table 2 and Figure 5 display the data.

Table 2. Data from Kristen Gilbert trial

	Gilbert working on shift	Gilbert not working on shift	Total
Death occurred on shift	40	34	74
Death did not occur on shift	217	1350	1567
Total	257	1384	1641

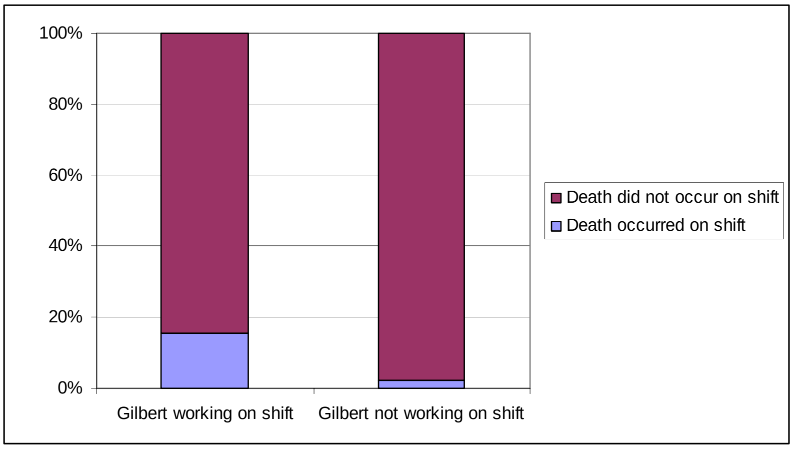


Figure 5. Results from Kristen Gilbert trial

As with the previous two examples, this study involves comparing two groups, with data presented in a 2×2 table, and investigating a conjecture that one group would "do better" on the response than the other. But a big difference is that, unlike the dolphin therapy experiment, this is an observational study in which no random assignment occurred. With this lack of randomness in the data production phase, some statisticians would argue that no randomization test should be performed here. But others would contend that we can still conduct a randomization test to assess whether the observed difference in death proportions (0.156 on a Gilbert shift vs. 0.025 otherwise) is large enough to infer that random variation is not a reasonable explanation. The p-value turns out to be less than 1 in a trillion, which effectively rules out “luck of the draw” as an argument for Gilbert’s defense.

Because this is an observational study, however, we cannot conclude that Gilbert’s presence on the shift is the *cause* of the higher death rate. The significance test does not rule out the possibility that other (confounding) variables may have differed between Gilbert shifts and no-Gilbert shifts. For example, without examining more detailed data, it is possible that Gilbert might have worked shifts during a particular time of day when deaths were more likely to occur.

This distinction in scope of conclusions that can be drawn from randomized experiments as opposed to observational studies is a key aspect of inference that I expect all students to learn. The GAISE guidelines for the K-12 curriculum include this distinction for secondary students.

5. SOME CONSIDERATIONS FROM PSYCHOLOGY

Why is this reasoning process difficult for many people (including, of course, students)? Part of the answer surely rests in all of the research that has shown how difficult probabilistic reasoning is for people (Nickerson, 2004). I particularly like how Keith Stanovich phrases this issue in his book *How to Think Straight About Psychology* (2007). Stanovich refers to probabilistic reasoning as the “Achilles heel of human cognition.” One example is that the concept of a statistical *tendency* is much more difficult for people to grasp than a deterministic relationship. Also, human beings are generally not comfortable with “luck of the draw” as an explanation; we tend to ascribe deterministic explanations to chance phenomena and tend not to consider variability in general, and chance variation in particular.

Notice also that the reasoning process of Fisherian inductive inference is related to a modus tollens argument in logic, but with a probabilistic aspect thrown in for good measure. Rethinking the “7/11 dice” example in these terms, we reason as follows:

1. If the dice were fair, it would be extremely surprising to observe a long string of sevens and elevens.
2. We observe a long string of sevens and elevens.
3. Therefore, we have extremely strong evidence that the dice are not fair.

In this particular context students seem to apply the reasoning process effectively and intuitively. But they often struggle to apply the reasoning process in less familiar contexts, and they often struggle mightily to understand it in the abstract. This bears many similarities to the well-known logic problem known as the Wason selection task (Wason & Johnson Laird, 1972).

The abstract version of the Wason task presents subjects with four cards that have a letter on one side and a number on the other. Subjects are then told the following rule: Every card with a vowel on one side has an even number on the other side. The four cards shown reveal an A, a B, a six, and a seven. Subjects are then asked which cards should be turned over in order to detect whether the rule has been violated. Studies show that a small percentage of people choose the correct answer, which is to turn over the A and the seven cards (Wason & Johnson Laird, 1972). Most people select the six rather than the seven card, failing to realize that the rule will be violated if the seven reveals a vowel on the other side (by a modus tollens argument). But when this same task is presented in concrete terms that are familiar to the subject, most people do indeed answer correctly. For example, suppose the rule is that all people who drink alcohol must be at least 21 years old. Consider four people: a 30-year-old, an 18-year-old, a beer drinker, and a soda drinker. Most subjects have no difficulty in realizing that they should check on how old the beer drinker is and on what the 18-year-old is drinking. The logical structure of this problem is identical to that with the cards, but the concreteness and familiarity make it much easier to solve correctly.

One of my points in mentioning the Wason task is that a modus tollens argument is hard for people to make intuitively, so the reasoning of statistical significance, which invokes a modus tollens-like argument with uncertainty thrown in, is all the more daunting. But my larger point is that students can apply this reasoning process for themselves if we start with concrete examples in a familiar setting.

6. COBB'S 3Rs

George Cobb (2007) refers to the randomization test approach to statistical inference as the 3Rs:

- Randomize data production.
- Repeat by simulation to see what's typical (and what's not).
- Reject any model that puts your data in its tail.

Cobb argues that introductory statistics students have a better chance of understanding the core logic of inference if it is presented in this manner as opposed to a more conventional approach based on calculations from normal-based probability distributions. Cobb writes: "Our curriculum is needlessly complicated because we put the normal distribution, as an approximate sampling distribution for the mean, at the center of our curriculum, instead of putting the *core logic of inference* at the center." While Cobb is referring to the introductory curriculum at the tertiary level, Scheaffer and Tabor (2008) advocate teaching statistical inference at the secondary level through this process of simulating randomization tests. Chance and Rossman (2006) adopt this approach in a tertiary course for mathematically inclined students.

The three examples discussed above have all involved categorical variables, but this 3Rs approach to statistical inference can also be applied with a quantitative response variable. Scheaffer and Tabor (2008) include such an example, as do Ernst (2004) and Cobb (2007). I prefer giving students experience with a categorical response variable first, because the quantitative response involves several complicating factors. One is that summarizing the difference between the groups is less straightforward; for example, you could use the difference in group means, or the difference in group medians, or some other statistic. Another complication is summarized by Wild (2006), who writes: "Assessment of 'significance' balances three factors—effect size, variability and sample size—in a very complicated way." By starting with categorical variables, we eliminate the variability issue because effect size and sample size are the only relevant factors.

7. POINTS OF CONTENTION, ALTERNATIVE APPROACHES

I should admit that the examples above involve some thorny issues on which statisticians disagree. In Example 2 about dolphin therapy, one question is why keep both margins (not only the number of subjects in each group but also the number who improved and did not improve) fixed when conducting the simulation. Lehmann (1993) points out that, although Fisher favored this analysis, others have criticized this procedure for being too conservative and having low power. Another tricky question is why to calculate the p-value by considering results *more extreme* than the actual results when calculating the p-value. The common answer is that with a large sample size, any one particular outcome is bound to have a small probability. (Imagine flipping a fair coin 10,000 times; obtaining 5000 heads is the most likely result but has probability only 0.008.) Bayesian statisticians do not accept this practice, however; a Bayesian analysis (discussed below) conditions on the data observed, not on data that were not observed.

The examples above all illustrate the Fisherian approach to statistical inference (1925, 1935a, 1935b). Fisher's approach emphasizes the strength of evidence provided by observed data against a null model. This strength of evidence is captured in the p-value, which measures the probability of having obtained such an extreme result (or more extreme) if the null model were true.

An alternative approach associated with Neyman (1935, 1955) adopts a more mathematical viewpoint. This view regards statistical inference as principally concerned with making a decision between competing hypotheses. The decision procedure is chosen optimally by specifying some condition on the two error probabilities. Typically this condition is to set the desired probability of type I error (rejecting a true null hypothesis), universally denoted by α.

Table 3 summarizes the different perspectives and emphases of the Fisher and Neyman approaches to statistical inference. In the last row of this table I suggest that the Fisherian approach is closer to informal inference and Neyman's to formal inference. As my earlier examples attest, I favor introducing students to the Fisherian approach.

Table 3. Comparing Fisher's and Neyman's perspectives on statistical testing

Fisher	Neyman
Significance testing	Hypothesis testing
Null model	Competing hypotheses
Strength of evidence	Error probabilities
Inductive inference	Inductive decision
p-value	α-level
Data-based	Mathematics-based
Informal inference?	Formal inference?

Are there implications of this contention for introductory teaching and learning? Hubbard and Bayarri (2003) contend that many teachers, authors, and researchers do not recognize and appreciate the differences between these approaches. They write: "Because statistics textbooks tend to anonymously cobble together elements from both schools of thought, however, confusion over the reporting and interpretation of statistical tests is inevitable." In his discussion of this article, Carlton (2003) argues that students "can handle both approaches" and suggests that α levels be introduced as prespecified thresholds for determining whether a p-value is small enough to constitute convincing evidence against the null model. Lehmann (1993) offers that the Fisher and Neyman perspectives are more compatible than others have realized. See Salsburg (2001) and Lehmann (2008) for readable accounts of the Fisher-Neyman dispute.

A third perspective takes a very different approach to statistical inference. All of the examples and discussion above, including both the Fisher and Neyman perspectives, adopt the classical (sometimes called frequentist) approach to statistical inference. Adherents of the Bayesian viewpoint adopt a subjectivist view of probability as measuring personal degree of belief in the proposition being considered. In the 7/11 dice example, a Bayesian would start with a prior probability that the dice are fair, before they are even rolled. Then the Bayesian updates this probability as the dice rolls are observed, using Bayes' Theorem as the mechanism. The result then is a conditional probability, given the observed data, that the dice are fair.

For example, suppose that you start by believing very strongly that the dice are fair, let's say with a probability of 0.99 (and with a very small 0.01 prior probability that the dice produce only sevens and elevens). Then after five consecutive rolls resulting in seven or eleven, Bayes' Theorem calculates that the updated (conditional) probability that the dice are fair becomes 0.051. In other words, those five rolls serve to reduce your belief that the dice are fair from a prior probability of 0.99 to a new probability of 0.051. After eight rolls of seven or eleven, your probability that the dice are fair drops to 0.0006. These are based on starting with a very high prior probability (0.99) that the dice are fair. But if instead you start with only a 0.5 probability that the dice are fair (so prior to seeing any rolls you think it's equally likely that the dice are fair or not), then the updated probability that the dice are fair drops to 0.0005 after five rolls of seven or eleven, and this probability falls all the way to 0.0000006 after eight such rolls.

Bayesians contend that talking of the probability that the dice are fair is very natural and interesting, yet this probability is nonsensical to a classical statistician: The dice are either fair or not; there's nothing random about that, so the classical statistician cannot assign a probability to that proposition. The only probability that can be determined by a classical statistician is the probability of obtaining such extreme results if a pair of fair dice are rolled repeatedly.

Proponents of the Bayesian approach cite many advantages for it. One is that it seems to correspond with how people actually reason. Another is that it results in probability statements about the null model being tested and for the parameter being estimated. Instructors of introductory statistics cringe when students interpret a p-value as a probability that the null model is true, or interpret a confidence interval by saying that there is a 95% chance that the parameter is within the interval. These statements are not only wrong but nonsensical from a classical approach, but they are quite appropriate and accurate from a Bayesian perspective.

A third advantage is that in many cases there truly is prior information that is relevant to making an inference. For example, suppose I tell you that I observed 8 members of a profession and saw that 4 were men and 4 were women. What would you infer about the overall proportion of women in that profession, if I tell you nothing else? But then what if I tell you that I am talking about mechanical engineers—would this information, and your prior knowledge about the relatively small proportion of engineers who are women, affect your inference? Or what if you learn that the occupation in question is pilots, or flight attendants? I suspect that you would draw quite different inferences from the same data in these situations, and quite appropriately so, based on your prior knowledge, or at least impressions, of the proportion of women in those various professions.

Another point of contention has emerged in the past decade, with a growing movement arguing that significance testing has been overused and misused, often serving as a substitute for thoughtful analysis, particularly in the social sciences. Harlow, Mulaik, and Steiger (1997) edited a collection of essays with the provocative title *What if There Were No Significance Tests?* A new book by economists Ziliak and McCloskey (2008) has the even more provocative title *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Two of the principal complaints lodged against significance testing are that with a large enough sample size, nearly all null models are rejected, and statistical significance does not necessarily imply practical significance. These critics often recommend estimating effect sizes as a replacement for assessing statistical significance.

Even proponents of significance testing admit the importance of the other major concept of statistical inference: estimating with confidence. The next section addresses this aspect of statistical inference.

8. INFORMAL REASONING ABOUT INTERVAL ESTIMATION

Example 4: Kissing couples Most people are right-handed and even the right eye is dominant for most people. Molecular biologists have suggested that late-stage human embryos tend to turn their heads to the right. German bio-psychologist Onur Güntürkün (2003) conjectured that this tendency to turn to the right manifests itself in other ways as well, so he studied kissing couples to see if they tended to lean their heads to the right while kissing. He and his researchers observed couples in public places such as airports, train stations, beaches, and parks. They were careful not to include couples who were holding objects such as luggage that might have affected which direction they turned. For each couple observed, the researchers noted whether the couple leaned their heads to the right or to the left. Of the 124 couples observed, 80 leaned to the right. Does this sample provide evidence that more than half of all kissing couples lean to the right? In light of the sample data, what proportion of the population might lean to the right?

We'll treat this sample as if it were a random one from the population of all kissing couples. As with the helper/hinderer study, we'll simulate 1000 repetitions of 124 couples that are equally likely to lean right or left. Figure 6 displays the results of one such simulation.

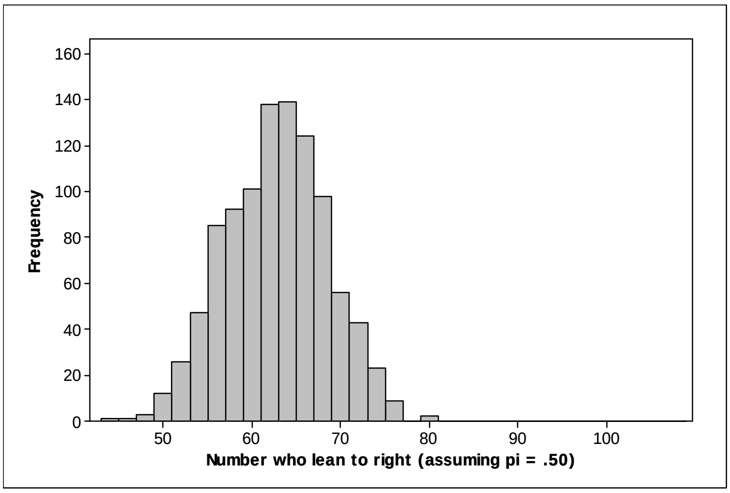


Figure 6. Simulation results for the kissing study

Notice that the observed result (80 couples who lean to the right) is way out in the tail of this empirical sampling distribution. The sample data therefore provide very strong evidence to reject that couples are equally likely to lean to the right or left. We have very strong evidence that kissing couples do indeed tend to lean to the right.

But the natural follow-up question is: How much more than half lean to the right? In other words, what proportion of the population of all kissing couples leans to the right? We can investigate the plausibility of values other than 0.5 by repeating the simulation analysis with those values. Our strategy remains the same: Reject any value of the population proportion that puts the observed data in the tail of its sampling distribution.

Figure 7 displays the results from six different simulations of 1000 repetitions each, changing the value of the population proportion each time.

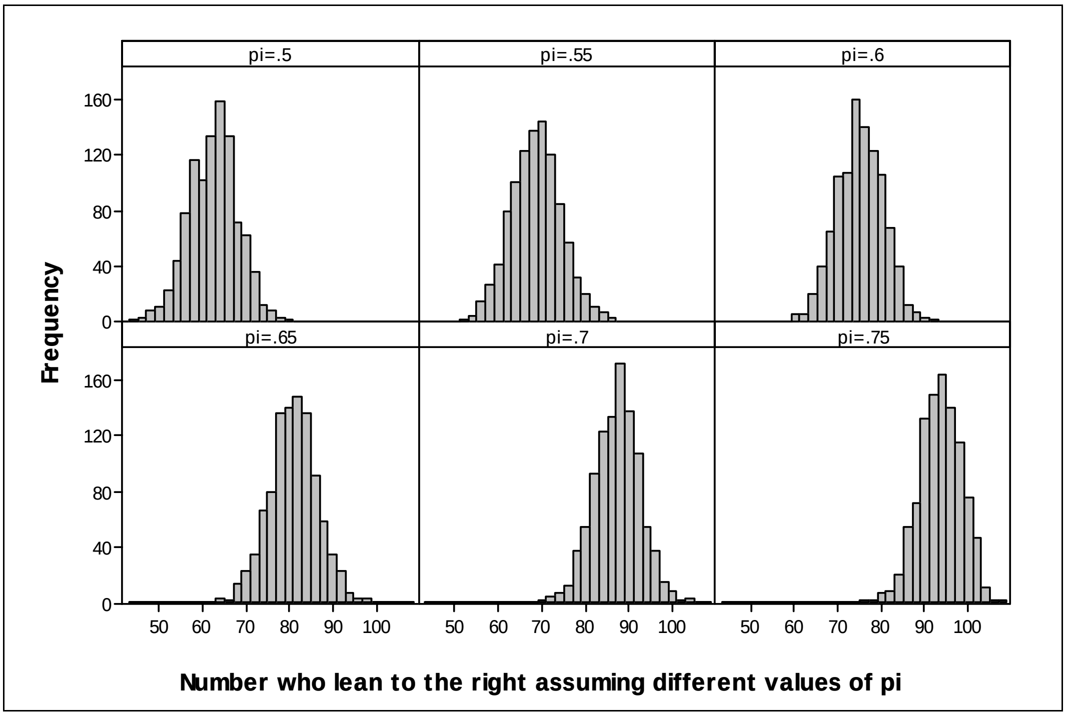


Figure 7. Comparing multiple models via simulation for the kissing study

Notice that the observed result (80 of 124 couples leaning to the right) is surprising (in the tail of the distribution) when the population proportion equals 0.5, 0.55, and 0.75, but not surprising when it equals 0.6, 0.65, and 0.7. Therefore, we include the values 0.6, 0.65, and 0.7 as plausible values of the population proportion who lean to the right. A more thorough analysis reveals an interval of plausible values (using a 5% level of significance) to be from about 0.56 to 0.73. This Fisherian approach to interval estimation is very different from calculating a confidence interval with a formula based on the normal distribution, such as: $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$. This simulation approach strikes me as a more informal method that is likely to involve and increase students' reasoning abilities.

9. CONCLUSION

I suggest that simulation of randomization tests provides an informal and effective way to introduce students to the logic of statistical inference. One advantage of this strategy is that it emphasizes the key role played by chance variation in statistical inference. During the SRTL-5 conference, Tim Erickson observed that asking this "what could have happened if the experiment/sampling had been repeated?" question is paramount in statistical inference. Harradine (2008) provided similar ideas and activities for introducing students to this issue. As the GAISE report suggests, understanding this reasoning process should be attainable by students at the secondary as well as tertiary levels.

I have emphasized categorical variables in the examples, in part because I think such variables provide a simpler context in which students can focus on key ideas of inference. I also suggest that the secondary curriculum often underutilizes categorical variables. I also worry that the secondary curriculum pays far more attention to sampling contexts rather than experimental studies, and I propose that more should be done with experimental studies and activities at this level.

ACKNOWLEDGEMENTS

I thank Joan Garfield and Dani Ben-Zvi for their vision in organizing the SRTL forums and for inviting me to speak to the group. Thanks also to Janet Ainley and David Pratt for providing such a hospitable environment for a very thought-provoking conference, as well as for editing this special issue. More thanks to all of the SRTL participants for their insightful presentations and discussions. Special thanks to Beth Chance for extensive and very helpful suggestions on early drafts of my remarks.

Many of the ideas presented here have emerged from a curriculum development grant from the National Science Foundation (NSF-DUE-CCLI #0633349), through very productive collaborations with Beth Chance, George Cobb, John Holcomb, and Matt Carlton.

REFERENCES

Antonioli, C., & Reveley, M. (2005). Randomized controlled trial of animal facilitated therapy with dolphins in the treatment of depression. *British Medical Journal, 331*(7527), 1231-1234.

Carlton, M. (2003). Comment on “Confusion over measures of evidence (*p*’s) versus errors (α’s) in classical statistical testing.” *The American Statistician*, *57*, 179-181.

Chance, B., & Rossman, A. (2006). *Investigating statistical concepts, applications, and methods*. Belmont, CA: Cengage.

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1*(1).
[Online: http://repositories.cdlib.org/uclastat/cts/tise/]

Cobb, G., & Gelbach, S. (2005). Statistics in the courtroom. In R. Peck, et al. (Eds.), *Statistics: A guide to the unknown*. Belmont, CA: Thomson.

Ernst, M. (2004). Permutation methods: A basis for exact inference. *Statistical Science*, *19*, 676-685.

Fisher, R. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Fisher, R. (1935a). The logic of inductive inference. *Journal of the Royal Statistical Society, 98*, 39-54.

Fisher, R. (1935b). Statistical tests. *Nature, 136*, 474.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., et al. (2005). *Guidelines for assessment and instruction in statistics education (GAISE) report: A Pre-K-12 curriculum framework*. Alexandria, VA: American Statistical Association.
[Online: http://www.amstat.org/education/gaise/]

Güntürkün, O. (2003). Adult persistence of head-turning asymmetry. *Nature, 421*, 711.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*, 557-560.

Harlow, L., Mulaik, S., & Steiger, J. (Eds.) (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Harradine, A. (2008, July). *Birthing big ideas in the minds of babes*. Paper presented at the IASE/ICMI Roundtable Conference, Monterrey, Mexico.

Hubbard, R., & Bayarri, M. (2003). Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing. *The American Statistician, 57*, 171-178.

Lehmann, E. (1994). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association, 88*, 1242-1249.

Lehmann, E. (2008). *Reminiscences of a statistician: The company I kept*. New York: Springer.

Neyman, J. (1935). Discussion of "Logic of inductive inference." *Journal of the Royal Statistical Society, 98*, 74-75.

Neyman, J. (1955). The problem of inductive inference. *Communications in Pure and Applied Mathematics, 8*, 13-46.

Nickerson, R. (2004). *Cognition and chance: The psychology of probabilistic reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ramsey, F., & Schafer, D. (2002). *The statistical sleuth: A course in methods of data analysis* (2nd ed.). Belmont, CA: Duxbury Press.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: W. H. Freeman.

Scheaffer, R., & Tabor, J. (2008). Statistics in the high school mathematics curriculum: Building sound reasoning under uncertainty. *Mathematics Teacher*, *102*(1), 56.

Stanovich, K. (2007). *How to think straight about psychology* (8th ed.). Upper Saddle River, NJ: Allyn & Bacon.

Wason, P., & Johnson Laird, P. (1972). *Psychology of reasoning: Structure and content*. London: Batsford.

Wild, C. (2006). The concept of distribution. *Statistics Education Research Journal*, *5*(2), 10-26.
[Online: www.stat.auckland.ac.nz/~iase/serj/SERJ5(2)_Wild.pdf]

Ziliak, S., & McCloskey, D. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.

ALLAN J. ROSSMAN
Department of Statistics
Cal Poly
San Luis Obispo, CA 93407

STATISTICAL COGNITION: TOWARDS EVIDENCE-BASED PRACTICE IN STATISTICS AND STATISTICS EDUCATION

RUTH BEYTH-MAROM
Department of Education and Psychology, The Open University, Israel
ruthbm@openu.ac.il

FIONA FIDLER
School of Psychological Science, La Trobe University, Melbourne, Australia
f.fidler@latrobe.edu.au

GEOFF CUMMING
School of Psychological Science, La Trobe University, Melbourne, Australia
g.cumming@latrobe.edu.au

ABSTRACT

Practitioners and teachers should be able to justify their chosen techniques by taking into account research results: This is evidence-based practice (EBP). We argue that, specifically, statistical practice and statistics education should be guided by evidence, and we propose statistical cognition (SC) as an integration of theory, research, and application to support EBP. SC is an interdisciplinary research field, and a way of thinking. We identify three facets of SC—normative, descriptive, and prescriptive—and discuss their mutual influences. Unfortunately, the three components are studied by somewhat separate groups of scholars, who publish in different journals. These separations impede the implementation of EBP. SC, however, integrates the facets and provides a basis for EBP in statistical practice and education.

Keywords: *Statistics education research; Statistical cognition; Statistical reasoning*

1. BACKGROUND

A wide range of research is relevant for improving statistical practice and statistics education, but we worry that this research is too fragmented for most effective use. We identify three facets of this research, and propose that the concept of *statistical cognition* can help bring these together, and provide a stronger basis for evidence-based practice (EBP) in statistics and statistics education. As an introductory example, consider confidence intervals (CIs), and three lines of discussion involving them.

First, for almost a century, mathematical statisticians have been studying CIs—developing theory and new applications, investigating robustness, and making comparisons with other inferential techniques. Second, within statistics, and in research fields that use statistics, there has been some discussion about possible misunderstandings of CIs; textbook authors also consider how to explain CIs, and possible misconceptions. However there has been almost no empirical study of how students and researchers think about CIs, or about misconceptions they may have. Third, there have been persistent calls for much wider use of CIs, in preference to null hypothesis significance testing (NHST), in psychology and other disciplines (e.g., Wilkinson et al., 1999). Reformers have

Statistics Education Research Journal, 7(2), 20-39, http://www.stat.auckland.ac.nz/serj

claimed CIs lead to better research decision making than NHST, and that students can more easily and successfully learn about CIs than NHST (e.g., Schmidt & Hunter, 1997). Our worry is that the evidence base especially for the second and third lines of discussion is sadly deficient, and that the three lines are not sufficiently integrated.

The first discussion, or facet, we referred to above was *normative*: theory of CIs, and techniques for their application, developed within mathematical statistics. The second considered how researchers, students, or others think about CIs—their informal statistical reasoning. This is the *descriptive* facet, which focuses on the cognition of using or teaching statistics. The third was the recommendation to replace NHST with CIs, and this is obviously *prescriptive*. The prescriptive facet seeks to improve statistical practice, and statistics learning. It might, for example, provide evidence about which CI diagrams and explanations are most effective in helping students achieve correct conceptualisations, as well as about which graphical designs and CI interpretations most successfully communicate research results.

By the 1980s the distinction between normative, descriptive and prescriptive was commonplace in judgment and decision making literature. “Decision Making: Descriptive, Normative and Prescriptive Interactions” was the name of a conference held in Boston at the Harvard Business School in 1983, the product of which was an edited book with this title (Bell, Raiffa, & Tversky, 1988). Those authors suggested the following taxonomy:

Descriptive:	(1) Decisions people make; (2) How people decide.
Normative:	(1) Logically consistent decision procedures; (2) How people should decide.
Prescriptive:	(1) How to help people to make good decisions; (2) How to train people to make better decisions. (p. 1-2)

For the purpose of the current discussion, we could substitute “statistical inferences” for “decisions.” We are interested in the mutual influences and contributions of these three facets, as well as their integration. One motivation for integration is to provide a more cohesive and complete evidence base for statistical practice and education.

Evidence-based practice (EBP) has a long history in medical decision making. The Institute of Medicine (2001) defined EBP as “the integration of best research evidence with clinical expertise and patient values” (p. 147). Psychology, nursing, social work, and other professional disciplines are progressively advocating and adopting EBP (Trinder & Reynolds, 2000). *Evidence-Based Medicine*, *Evidence-Based Child Health*, *Evidence-Based Communication Assessment and Intervention*, *Evidence-Based Complementary and Alternative Medicine*, *Evidence-Based Library and Information Practice* are all relatively new journals aimed to alert professionals to important theoretical and empirical advances in their profession that might contribute to improved decision making in their professional practice. Similarly, a desire to ensure that students meet high standards has increased the demand for EBP in education (Davies, 1999). Statisticians and statistics educators should likewise adopt EBP by, wherever possible, using relevant evidence from research to guide what they do.

Within medical EBP, successful implementation of research into practice requires integration of three core elements: relevant *evidence*, the context or *environment* into which the research is to be placed, and the *method* or way in which the process is accomplished (Kitson, Harvey, & McCormack, 1998). There is some correspondence between these elements and our normative, descriptive and prescriptive facets, respectively. If a statistician is advising a researcher about data analysis for a report, normative information about statistics provides the *evidence*, for example, statistical theory about correlation. Descriptive information about likely misunderstandings of correlation by the readers of a journal article is part of the researcher's *context*; and prescriptive information—if available—suggests how most effectively to present correlations and thus provides a *method*.

We therefore believe that these three lines of research are necessary to build an evidence base for statistical practice and education, and that adoption of EBP directly depends on the integration of these fragmented research facets. In this article, we first introduce statistical cognition—as a concept and an integrative field—in further detail. Second, we explore the interactions among the normative, descriptive, and prescriptive facets. Some of these interactions may be obvious, but others are subtle and still others virtually missing. Exploring these relationships helps identify gaps in current research, and priorities for future research. Third, we describe two examples to illustrate these interactions in statistics teaching and practice. We then briefly examine institutional and sociological factors that have contributed to the fragmentation of the normative, descriptive, and prescriptive facets of research. Finally, we explore how statistical cognition may overcome some of the barriers that currently impede integration, and make recommendations about how this integrated field should proceed.

2. STATISTICAL COGNITION

Cognition is usually defined as the mental processes, representations, and activities involved in the acquisition and use of knowledge. Statistical cognition is accordingly defined as the processes, representations, and activities involved in acquiring and using *statistical* knowledge. What are the issues relevant in the study of statistical cognition? One aspect is how people acquire and use statistical knowledge and how they think about statistical concepts—this is the descriptive facet of statistical cognition. The study of how people *should* think about statistical concepts—the normative—is also an important aspect of statistical cognition as this is often what we are exposed to (e.g., in school) and it is also the standard to which our performance is usually compared. Finally, the question of closing the gap between the descriptive (the "is") and the normative (the "should")—the prescriptive—is a critical issue in statistical cognition.

As such, statistical cognition is a field of theory research and application concerned with normative, descriptive, and prescriptive aspects. It focuses on (a) developing and refining normative theories of statistics and their application, (b) developing and testing theories explaining human thinking about and judgment in statistical tasks, and (c) developing and testing pedagogical tools and ways of communication for the benefit of practitioners and teachers.

Statistical reasoning, a term already widely used (Garfield, 2002; Garfield & Gal, 1999), concerns the mental processes which shape the process and representations of statistical cognition. As such, it is concerned mainly with the descriptive facet. However, statistical cognition, like mathematical cognition, takes a broader approach encompassing normative and prescriptive research, in addition to the descriptive research found in the literature on statistical reasoning and in the experimental and educational psychology literatures. Statistical cognition therefore integrates the three lines of research we believe are needed for effective EBP.

3. THE THREE FACETS OF STATISTICAL COGNITION: NORMATIVE, DESCRIPTIVE AND PRESCRIPTIVE

The science of statistics contributes most to the normative facet of statistical cognition. It includes simple rules (e.g., the conjunction rule of probability), theorems and laws (e.g., Bayes' theorem, the law of large numbers), as well as models (e.g., for estimation and inference). Statisticians may agree that a particular normative solution to a problem is best, or may hold differing views as to which normative model should be applied. Both Frequentist and Bayesian approaches have been developed and advocated within the normative facet, as has theory for both NHST and estimation. Whether consensus or controversy dominates, such rules, models, and approaches comprise the normative facet of statistical cognition.

Dissemination of statistical information (beginning in the 19th century) about many aspects of society has increased the need for laypersons as well as professionals to understand statistical concepts. Many reports in the mass media (about psychological, medical, economic, or political issues) can only be correctly comprehended with an understanding of statistics. As early as the beginning of the 20th century, H. G. Wells emphasised the importance of teaching statistical reasoning to produce an educated citizen, with statistical reasoning being as important as reading and writing (Huff, 1973, p. 6). In the early 1980s the concept of 'statistical numeracy' was first introduced as a sub-category of 'numeracy':

> Statistical numeracy requires a feel for numbers, and appreciation of appropriate levels of accuracy, the making of sensible estimates, a commonsense approach to the use of data in supporting an argument, the awareness of variety of interpretation of figures, and a judicious understanding of widely used concepts such as means and percentages. (Cockcroft, 1982, paragraph 781)

The broader term 'statistical literacy' (Ben-Zvi & Garfield, 2004; Gal, 2002; Wallman, 1993) later replaced statistical numeracy, and became an important goal yet to be achieved. The need to enhance statistical literacy has been gradually recognised with the publication of psychological research assessing intuitive statistical reasoning (e.g., Edwards, 1968; Meehl, 1954; Tversky & Kahneman, 1974) and studying the cognitive processes involved (e.g., Gilovich, Griffin, & Kahneman, 2002; Kahneman, Slovic, & Tversky, 1982; Sedlmeier, 1999). These developments shaped two lines of theory, research, and applications: the descriptive and the prescriptive approaches.

"Man as an intuitive statistician" (Peterson & Beach, 1967) was the first comprehensive publication on intuitive statistical reasoning and it opened a long-lasting debate about lay persons' as well as experts' capabilities. Tversky and Kahneman's (1974) seminal work replaced Peterson and Beach's optimistic view with the heuristic and biases model: Intuitive statistical judgments are often based on a limited number of simplifying heuristics rather than on more formal and extensive algorithmic processing. These heuristics can give rise to systematic errors, or biases. These lines of research—the evaluation of people's statistical reasoning and the cognitive processes underlying them—are the core of the descriptive aspect of statistical cognition.

Statistical education aims to improve statistical reasoning. The best approaches and tools for reaching this goal, and the pedagogical prescriptions for the teaching of statistics, should be based on the art, science, and profession of teaching. Learning by doing (e.g., Glaser & Bassok, 1989; Smith, 1998), authentic learning (e.g., Donovan, Bransford, & Pellegrino, 1999; Mehlinger, 1995) and situated cognition (e.g., Brown, Collins, & Duguid, 1989) are examples of educational or instructional theories that have direct pedagogical recommendations.

A statistical consultant may advise that a particular model and statistical analysis is appropriate for the data of interest—relying on normative considerations. The question then becomes how the results will be written up for publication, and that is a question of statistical communication: What numerical, graphical or other information should be presented so that target readers will understand most accurately what was found and what conclusions are justified? Those questions should be in the forefront of the mind of the statistical consultant, as well as the researcher, and it is the job—we would argue—of statistical cognition to provide research-based guidance as to how statistical communication can be best accomplished. Similarly, the discipline of statistics provides much content for the statistics curriculum, but it is the job of statistical cognition to provide guidance for teachers on how to best achieve accurate and appropriate statistical learning.

Each of the three approaches has theory, research, and applications rooted primarily in different disciplines (statistics, psychology, education). As we have indicated, we believe there has been insufficient interaction between them. We hope that statistical cognition can encourage closer collaboration among the approaches, and thus develop a body of research that can support EBP in statistics. This body of research should focus on projects like statistical reasoning of laypersons as well as experts; developmental aspect of statistical reasoning along the life span; cognitive, social and neurological processes that underlie statistical reasoning; and testing the efficiency of instructional techniques, approaches, and tools.

Figure 1 illustrates in a schematic way the three facets of statistical cognition, and the arrows indicate paths of influence. The normative facet (N) specifies what statistical techniques can correctly be applied in a given situation; it is potentially informed by the full body of knowledge that is mathematical statistics. The descriptive facet (D) comprises knowledge of how people think about statistical concepts, what messages they receive when inspecting a statistical presentation, and their statistical misconceptions and biases. Psychology has provided most of the information in D, yet this information is scanty and there are many important gaps that need further research. The prescriptive facet (P) comprises knowledge about how to achieve successful statistical communication and education. This knowledge, such as it is, has largely come from psychology and education, and again much additional knowledge is needed in this facet. The contribution of the normative facet (N) to the prescriptive (P) is large and probably straightforward to grasp: It is probably most natural and common to base advice or teaching on statistical theory. There can perhaps (the dotted arrow) be influence in the reverse direction, when experience with advising or teaching (that’s P) prompts development of additional theory (N). The next sections will focus on the two-way influences between N and D, and between D and P. We consider both the known and the potential contributions relevant to each arrow.

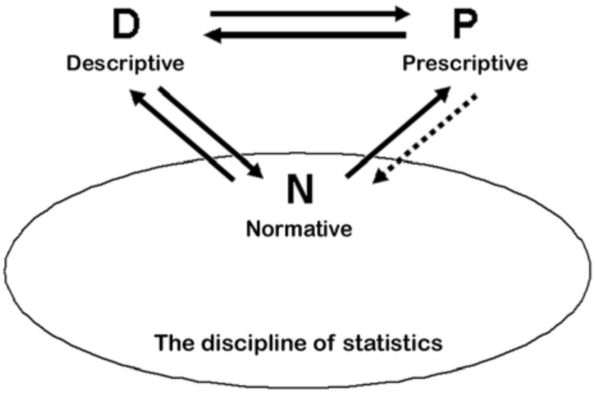


Figure 1. Schematic relations between the proposed three facets of statistical cognition

4. CONTRIBUTION OF THE NORMATIVE TO THE DESCRIPTIVE

The normative rules, theories, and models of the science of statistics are the standards recommended for summarizing data, interpreting it, and evaluating hypotheses. These are the norms used by professionals when analysing empirical research or advising researchers. However, these norms have also been used as standards to which intuitive statistical reasoning (of laypeople and experts) is compared. For example, people's performance in solving conjunction tasks has been compared to the predictions of the conjunction rule: $P(A\&B) \leq P(A)$ and $P(A\&B) \leq P(B)$ (Tversky & Kahneman, 1983). Normative standards have been used similarly in research on people's judgments of disjunctive probabilities (Bar-Hillel & Neter, 1993), conditional probabilities (Pollatsek, Well, Konold, Hardiman, & Cobb, 1987), effects of sample size (Bar-Hillel, 1979), judgment of randomness (Falk & Konold, 1994, 1997), interpreting p-values in hypothesis testing (Falk, 1986; Oakes, 1986)—to mention but a few cases.

A normative model can thus provide a theoretical framework for describing how people should perform a task. It can also identify a set of logically possible deviations from the model, which can be tested empirically. Such an approach was used by Fischhoff and Beyth-Marom (1983). They adopted Bayesian inference as a general framework for characterizing people's hypothesis evaluation behaviour in terms of its consistency with or departures from the model. They identified the kinds of systematic deviations from the Bayesian model that could, in principle, be observed, and presented evidence demonstrating their actual existence. Normative models provide a reference for the evaluation of people's performance in statistical tasks (the descriptive facet).

The choice of the appropriate normative model may seem obvious, but sometimes is debatable, or may be thrown into doubt after further consideration of descriptive results. Gigerenzer (1991), for example, argued that probability theory is imposed as a norm for judgments about a single event in research on the conjunction fallacy, and this would be considered misguided by statisticians who hold that probability theory is about repeated events. A further example is Cohen's (1979) questioning of the choice of a Bayesian model as a normative standard in Tversky and Kahneman's (1974) descriptive work; he suggested an alternative normative Baconian model. Thus, the choice of a normative standard to which people's performance is compared must be made with much care, being sure that the assumptions underlying the normative model (e.g., random sampling), are also part of the judgmental task performed by people.

5. CONTRIBUTION OF THE DESCRIPTIVE TO THE NORMATIVE

How can judgmental tasks, and people's performance of them, contribute to the relevant normative model? The historical account of NHST, empirical research on the understanding of p-values and, more generally, of people's intuitive inferential reasoning, provides one example of such a contribution.

NHST in its contemporary form (a hybrid of two schools of thought, one associated with Fisher, the other with Neyman and Pearson) was gradually applied in empirical research from 1940 (Hubbard & Ryan, 2000). There has been controversy about NHST since its inception, and the number of published works critical of it has increased dramatically since then (Anderson, Burnham, & Thompson, 2000).

The most common arguments against NHST refer to a catalogue of misconceptions about p-values. This catalogue (which is descriptive) has been built over many years from teachers' observations (e.g., Schmidt & Hunter, 1997), surveys of journal reporting practices (e.g., Finch, Cumming, and Thomason, 2001; Fidler et al., 2005) and empirical studies with researchers and students (Haller & Krauss, 2002; Kalinowski, Fidler, &

Cumming, 2008; Oakes, 1986). That the misconceptions are widespread and robust is well known and often demonstrated. They have also been compiled and summarised often. Kline (2004), for example, listed five common fallacies in the interpretation of p-values and eight common fallacies in reaching conclusions after deciding to reject or failing to reject the null hypothesis based on a p-value.

There are certainly advocates of statistical reform who believe that such misconceptions are the overwhelming, if not sole, problem with NHST. Rossi (1997), for example, stated "whereas some see significance testing as inherently flawed, I believe the problem is better characterised as the misuse of significance testing" (p. 175). However, there are others who hold the position that, even if used and interpreted properly, NHST contributes little knowledge and "is not the way any [proper] science is done" (Cohen, 1994, p. 999). A stronger expression of this position is that the procedure is itself fundamentally flawed; that NHST has a "flawed logical structure" (Falk & Greenbaum, 1995, p. 75).

There is also a third position, which draws the previous two together, and illustrates how the descriptive can contribute to the normative. This is the position that NHST is so widely misinterpreted *precisely because* the underlying logic is flawed. As Kline (2004) explained, "false beliefs may not be solely the fault of the users of statistical tests. … This is because the logical underpinnings of contemporary NHST are not entirely consistent" (p. 9). Kline is referring to the conflicting Fisherian and Neyman-Pearsonion paradigms that have become the institutionalised hybrid of NHST. Schmidt and Hunter (1997) provided another illustration of how descriptive considerations have challenged the normative status of NHST: "Any teacher of statistics knows that it is much easier for students to understand point estimates and CIs than significance testing with its strangely inverted logic" (p. 56). For these critics, challenges to the normative status of NHST have (at least in part) emerged from descriptive work on misconceptions.

Another alternative to NHST is Bayesian hypothesis testing, which differs from NHST in its interpretation of probability, and on these three principles: (1) Prior probabilities have to be taken into account; (2) alternative hypotheses play a role in the testing of a null hypothesis; and (3) the focus of analysis is P(H|D), and not P(D|H), where H is a hypothesis and D some data. We believe Bayesian methods have not been widely accepted at least in part because of users' misconceptions. That is, that their normative status has been in part determined by obstacles that are descriptive in nature. Research on the base rate fallacy (Bar-Hillel, 1980) demonstrated how people tend to ignore base-rates, thus behaving like null hypothesis statistical testers. Research on pseudo-diagnosticity (Beyth-Marom, 1990; Beyth-Marom & Fischhoff, 1983) indicated that people often base their updating of a hypothesis on the magnitude of P(D|H), ignoring P(D|~H), thus again behaving like null hypothesis statistical testers. There is also overwhelming evidence that people often confuse P(H|D) and P(D|H) and use, incorrectly, P(D|H) as their estimate of P(H|D) (Eddy, 1982; Haller & Krauss, 2002; Oakes, 1986). Thus, although people demonstrate severe misconceptions of the NHST model, by ignoring base rates and the relevance of alternative hypotheses, and by using P(D|H) for P(H|D), their intuitions remain more in line with the NHST model than the Bayesian one.

Descriptive findings also shed light on the history and development of normative models in science more broadly. Research on the perception of different scientific concepts (e.g., in physics, mathematics, and biology) by laypersons and experts has repeatedly shown that intuitive concepts deviate systematically from normative ones. Often the intuitive beliefs were similar to earlier, and now discredited, scientific theories. Erickson (1980), for example, investigated the change of children's viewpoints about heat from a Caloric viewpoint to a Kinetic one. This developmental change has its counterpart in the evolution of physics. Perhaps a parallel in statistical inference is yet to occur: Researchers' intuitions, and their statistical practices, are both still largely at the NHST stage; advancement to, for example, Bayesian thinking, and Bayesian techniques, is still largely for the future. (We recognise that development of Bayesian methods pre-dated development of NHST; it is widespread adoption that is the focus here.)

Naïve statistical concepts can thus influence the normative theories of statistics; first, by offering insight into their evolution and, second, by questioning their validity and contributing to their development and change.

6. CONTRIBUTION OF THE DESCRIPTIVE TO THE PRESCRIPTIVE

The idea that descriptive should influence prescriptive may be familiar, but we believe that fully exploiting the potential contribution of the descriptive facet to the prescriptive is the most serious challenge in the triangle of Figure 1. Many practitioners and teachers of statistics, and authors of statistical textbooks, are only vaguely aware of the substantial cognitive literature on statistical reasoning and the contributions it can make.

Students young or old don't enter the learning arena 'tabula rasa' (Pinker, 2002), but already holding beliefs about scientific concepts and processes. They also have everyday meanings for words that are used in a more specialized way in science. These beliefs might help or hinder learning depending on their consistency or discrepancy with what is taught.

From this perspective, educators have been interested in students' 'preconceptions', 'naïve conceptions', or 'naïve theories'. If those were found to be inconsistent with formal concepts to be taught they were regarded as 'misconceptions' or 'alternative conceptions'. Misconceptions may come from strong word association, confusion, conflict, or lack of knowledge (Fisher, 1985). They usually share the following characteristics: (a) they are at variance with normative conceptions held by experts in the field; (b) they tend to be pervasive (shared by many different individuals), and (c) they are often highly resistant to change, at least by traditional teaching methods. Thus, special teaching methods have to be developed. For example, some educators recommend that teachers should be given numerous examples of how to identify misconceptions held by pupils and strategies to change them (Lawrenz, 1986; Smith & Anderson, 1984). Others have suggested starting the teaching with students' ideas and then devising teaching strategies to take some account of them (Engel Clough & Wood-Robinson, 1985).

Most research on naïve conceptions and misconceptions—the descriptive element of statistical cognition—originated in cognitive psychology: how people reason under uncertainty. As uncertainty and statistical information surrounds us, efficiently coping with it is essential for our everyday conduct. Moreover, statistical reasoning is a tool used by experts in carrying out and interpreting research. Thus teaching of (normative) statistics is essential in schools—for the developing of good statistical reasoning—and in university—for doing research and interpreting its results. However, there is evidence (e.g., Abelson, 1995; Sedlmeier & Gigerenzer, 1989), as well as widespread classroom experience, to suggest that the teaching of statistics is often not very successful. In a literature review of the teaching of statistical reasoning to students at college and precollege levels, Garfield and Ahlgren (1988) concluded, two decades ago, that "little seems to be known about how to teach probability and statistics effectively" (p. 45). More recent research has had some impact on college teaching, but many courses remain unaffected by its outcomes (Garfield, Hogg, Schau, & Whittinghill, 2002; Ben-Zvi & Garfield, 2004). In other sciences, at least some educators are aware of the relevance of misconceptions (the descriptive facet) to the effective teaching of science (the prescriptive facet), but it seems that statistical instructors are less often aware of students' statistical misconceptions, and have few instructional tools designed to overcome them. Cognitive psychology, however, can offer considerable research on naïve conceptions and misconceptions, and how people reason under uncertainty—the descriptive element of statistical cognition.

Descriptive research on informal statistical reasoning might contribute to statistics teaching not only by identifying misconceptions, but also by describing the processes underlying them. This research can guide the development of effective teaching strategies. In his book *Improving statistical reasoning: Theoretical models and practical implications* Sedlmeier (1999) identified four descriptive explanatory models of statistical reasoning, and derived from them implications for statistical teaching. According to Sedlmeier's 'adaptive algorithms' explanation, the human mind is equipped with evolutionarily acquired cognitive algorithms that are able to solve complicated statistical tasks. These algorithms work for frequencies, but not for probabilities or percentages. The instructional implication is that to improve performance we should teach people how to translate from the format given in the task (e.g., probabilities) into natural frequencies (Gigerenzer & Hoffrage, 1995).

The heuristics and biases cognitive psychology literature has adopted dual-process theory (Sloman, 1996; Stanovich & West, 2000), which identifies two quite different cognitive modes, System 1 (S1) and System 2 (S2), approximately corresponding to the common sense notions of intuitive and analytical thinking. The two systems differ in various ways, most notably on the dimension of accessibility: how fast and how easily things come to mind. Many of the non-normative answers people give to statistical (as well as other) questions can be explained by the quick and automatic responses of S1, and the frequent failure of S2 to intervene in its role as critic of S1. Based on this dual-process theory, Kahneman in his Nobel Prize lecture set an agenda for research:

> To understand judgment and choice we must study the determinants of high accessibility, the conditions under which S2 will override or correct S1, and the rules of these corrective operations. Much is known of each of the three questions, all of which are highly relevant to the teaching of statistical reasoning. (Kahneman, 2003, p. 716)

Leron and Hazzan (2006) demonstrated how dual-process theory and empirical results from heuristic and biases research might shed light on mathematics education. They argued that the most important educational implication is "to train people to be aware of the way S1 and S2 operate, and to include this awareness in their problem solving toolbox" (p. 123). Such a toolbox is relevant also for statistical reasoning.

Communication of statistical information in newspapers and magazines, as well as in statistical textbooks and courses, includes many words used for statistical concepts that are also used in common language, such as 'or', 'chance', 'randomness', 'confidence', 'precision', and 'correlation'. Often, the meaning in everyday language is similar to the technical meaning in the science of statistics. However, sometimes the two concepts do not overlap. For example, Beyth-Marom (1982) showed how laypersons interpret correlation between two asymmetric variables, such as 'pneumonia: pneumonia vs. no pneumonia', where the values have differential status, and the variable has the name of one of its values. Participants interpret such correlations as the tendency of the two 'present' values to coexist, thus interpreting relationship between variables as a relationship between values, and ignoring all other information relevant for the evaluation of statistical correlation. This everyday interpretation of correlation is consistent with the *Oxford English Dictionary* (2007) definition of correlation as "a mutual relationship." This dictionary, as well as *The American Heritage* (2000) and *Webster's Online* (2007) dictionaries, present two or more definitions of correlation: one specific for statistics (mentioning variables) and the other the common everyday interpretation (mentioning entities or things). Falk and Konold (1994, 1997) showed a similar phenomenon in the perception of randomness.

When ordinary language is used for reasoning, conceptions and misconceptions are often shaped by the nature of the social interaction and the conversation taking place (Grice, 1975). "Respondents [students] deviate from the judgments predicted by the normative model considered relevant by the experimenter [teacher] by using rules of conversational inference very different than those assumed by the experimenter [teacher]" (Hilton, 1995, p. 266). Recognition of linguistic and conversational factors, as well as being alert to possible discrepancies between statistical and conversational meanings, is likely to have practical pedagogical implications for improving statistical understanding.

A further contribution of D to P is developmental research on statistical reasoning, beginning with Piaget and Inhelder (1975). The introduction of statistics into the school curriculum has prompted more attention to understanding developmental aspects of statistical literacy. A number of models of cognitive development in probability and statistics have been proposed (Biggs & Collis, 1991; Jones, Langrall, Thornton, & Mogill, 1997; Mooney, 2002; Watson & Callingham, 2003). According to these, instructional materials should be appropriate for students' age and cognitive development. For example, the model suggested by Jones and his colleagues (1997, 1999) includes four statistical concepts, the comprehension of which develops through four levels. Kafoussi (2004) used this model to guide the development of children's instructional activities, and then the analysis of their understanding of probability.

In our view, the substantial amount of cognitive knowledge on informal statistical reasoning has the potential to guide development of effective strategies for improving the statistical understanding of students and also researchers.

7. CONTRIBUTIONS OF THE PRESCRIPTIVE TO THE DESCRIPTIVE

We argued above that research on intuitive statistical reasoning (the descriptive facet) can guide instructional recommendations, and the design of teaching strategies, materials, and tools (the prescriptive facet). Evaluation of these prescriptive procedures gives information about their practical effectiveness, and also tests the underlying descriptive theory, enhancing or weakening its validity. Prescriptive research can thus contribute to the descriptive. For example, Sedlmeier's (1999) training program mentioned above demonstrated how the descriptive facet can influence the prescriptive. Sedlmeier used evaluation of his training programs (part of the prescriptive facet) to test the descriptive theories, thus demonstrating the interplay between the descriptive and prescriptive facets of statistical cognition. Statistical classrooms are the arena where the prescriptive is introduced, and where descriptive theories can be tested and be refined by evaluating the influence of different training programs.

8. INTEGRATION OF THE THREE FACETS IN STATISTICAL EDUCATION AND PRACTICE: TWO EXAMPLES

Statistics textbooks are based on the normative facet, but often incorporate also the author's descriptive and prescriptive ideas. Teachers and statistical consultants often use, in addition to the textbook and their statistical expertise, various pedagogical strategies to help students and researchers understand statistical concepts. They may consider the intuitive perceptions students and clients have at the start. They thus call on their own descriptive and prescriptive ideas in their efforts to assist the students and researchers achieve good understanding. Do these ideas reflect their clients' conceptions and misconceptions? Research evidence can shed light on the validity of these professionals' mis/conceptions.

8.1. CORRELATION BETWEEN TWO DICHOTOMOUS VARIABLES

Consider, as a first example, descriptive research on a basic topic: correlation between two dichotomous variables. We will mention the normative model, describe research results from the descriptive facet and discuss possible implications for teaching, thus illustrating the mutual influence of the three facets of statistical cognition, and the research needed for EBP.

Adults' perception of the correlation between dichotomous variables had been examined in a number of studies (Beyth-Marom, 1982; Jenkins & Ward, 1965; Shaklee & Tucker, 1980; Smedslund, 1963; Ward & Jenkins, 1965). Usually, participants were given pairs of values, and asked to estimate the direction and/or strength of the relationship between the variables. Most research evidence found estimates were biased relative to the normative measure, which is based on the difference between two conditional probabilities. Participants' estimates were a function of the way data were presented (trial by trial, as a list of data pairs, or in a summary table); of the instructions given; and of whether asymmetric or symmetric variables were used (Beyth-Marom, 1982).

Task instructions were varied because the experimenters recognised that technical and lay usage of correlation, and other terms, may be very different. The kind of explanation participants were given was found to influence the estimates given, thus indicating how sensitive people are to language usage.

The asymmetric-symmetric distinction refers to whether the two values of the variable were different or similar. In the asymmetric case (e.g., pneumonia vs. no pneumonia; symptom present vs. symptom absent) one value has a lower status than the other, whereas in the symmetric case (e.g., gender: male, female), the values have a similar status. In the asymmetric case, the name of the variable is like the name of one of its two values (variable 'pneumonia', values 'pneumonia' and 'no pneumonia'). With symmetric variables, the name of the variable ('gender') differs from the name of its two values ('male', 'female'). Furthermore, in the asymmetric case, the two values may be described as 'occurrence', 'non-occurrence'. A 'non-occurrence' or 'negative' event has less impact on people's attention than a positive event (Nisbett & Ross, 1980). Participants' perception of correlation was much more biased for asymmetric variables, for which they tended to perceive only one or two cells of the full 2×2 table that is required for the normative assessment of correlation. With symmetric variables they tended to take account of all four cells—although not necessarily using the correct formula.

This research on naïve perceptions of correlation has a number of pedagogical implications:

1. When trying to explain a statistical concept, like correlation, teachers as well as students have to be aware of any different connotations a term may have in day to day language and in statistics.
2. The comprehension of correlation depends on a clear perception of the difference between variables and values.
3. It may be better first to use symmetric variables and then, after students understand that all four cells are relevant for the assessment of correlation, present examples involving asymmetric variables. Discussion of those might highlight the importance of the alternative values ('no symptom', 'no pneumonia') in estimating correlations and in other statistical tasks; and

4. Table format should be used at first, because students understand this format best. Then, later, students can work with other data presentation formats, perhaps by creating the 2×2 table themselves.

Pedagogical implications such as these illustrate how D can make a valuable contribution to P. These implications should shape training; the training should then be evaluated, thereby testing the validity of students' intuitive conceptions and the validity of the pedagogical implications. This demonstrates the pathway by which P may influence D.

It is unfortunate, although valuable, that the examination of issues in terms of our three facets identifies gaps in research knowledge that is essential for the effective adoption of EBP.

8.2. CONFIDENCE INTERVALS

For our second example, CIs, we focus on statistical practice—in particular, the formulation of advice for researchers. We have already introduced this example, but here we offer a quick review of the three facets of CI research. Normative research includes CI theory in mathematical statistics, and presentations intended to help researchers use CIs for data analysis (e.g., Altman, Machin, Bryant, & Gardner, 2000). The descriptive facet includes study of how researchers understand and interpret CIs. Prescriptive research includes study of how best to improve statistical practice: It provides evidence about what graphical design for CI figures and what wording used to interpret CIs most successfully communicate research results.

Now consider what research is available on CIs. Normative information is abundant, and journals continue to publish further theoretical results and applied techniques. In stark contrast, there is very little descriptive evidence about people's CI thinking. Cumming, Williams, and Fidler (2004) found that many researchers in psychology, behavioural neuroscience, and medicine hold the misconception that a 95% CI is also a 95% prediction interval for a replication mean, whereas a 95% CI has an average 83% chance of including the mean of a replication experiment. Belia, Fidler, Williams, and Cumming (2005) reported evidence of further misconceptions widely held by researchers about 95% CIs. These are some of the very few examples of descriptive research on CIs.

The next step is to suggest improved guidance for researchers and better graphical conventions for presenting CIs in figures and study whether these improvements are effective in overcoming those misconceptions: That, of course, is prescriptive research.

Cumming and Finch (2005) described a number of *rules of eye*, intended as simple guidelines for interpretation of 95% CIs shown in figures. Use of these rules would overcome misconceptions identified by Belia et al. (2005). Cumming (2007) presented figures and simulations to illustrate the relation between CIs and p-values; these were also intended to overcome some of the problems identified by Belia et al. The rules of eye, and illustrations of how CIs and p-values relate, are within the prescriptive facet, but need to be evaluated and found effective to become part of that facet's contribution to the EBP of statistics.

Our CI example identifies the importance of all three facets and their interactions, and emphasises the deficiencies of current descriptive and prescriptive knowledge. Considering statistical practice, there is some descriptive research identifying problems, but almost no prescriptive research showing what changes in practice, or what guidance to researchers, can be effective in overcoming those problems. It is especially important to expand descriptive and prescriptive knowledge about CIs because, as we mentioned earlier, persisting criticism of NHST is leading to recommendations that CIs be much more widely used in psychology and other disciplines (Cumming & Fidler, in press;

Fidler & Cumming, 2008). This highly desirable reform of statistical practice is being hampered by lack of evidence about effective ways to overcome CI misconceptions.

9. BARRIERS TO INTEGRATION

Despite the mutual influence of the three facets of statistical cognition and the obvious need for integration, their current state of fragmentation is a major obstacle to building a cohesive evidence base for statistical practice and education. Below are some quotations that illustrate the fragmentation; they will be familiar to many readers.

1. "… the almost universal reliance on merely refuting the null hypothesis … is basically unsound, poor scientific strategy and one of the worst things that ever happened in the history of psychology" (Meehl, 1978, p. 817).
2. "The believer in the law of small numbers … rarely attributes a deviation of results from expectations to sampling variability, because he finds a causal 'explanation' for any discrepancy" (Tversky & Kahneman, 1982, p. 29).
3. "[Statistical] power is neglected by psychologists because, given their typically mistaken understanding of statistical significance, it is an unnecessary concept" (Oakes, 1986, p. 83).
4. "Activities specifically designed to help develop students' statistical reasoning should be carefully integrated into statistics courses" (Garfield, 2002).
5. "Since it appears that in judging randomness, subjects attend to the complexity of sequences, it might be possible to foster a more intuitive, yet mathematically sound, conception of randomness if it is introduced via the complexity interpretation" (Falk & Konold, 1994, p. 10).

The first quotation is an example of the vast literature that advocates reform of statistical inferences practices, questioning the normative justification of NHST. The next two describe people's intuitive perceptions (laypersons' as well as experts') of three statistical concepts: sampling, statistical power, and statistical significance. The fourth makes a prescriptive recommendation about the teaching of statistics. The final quotation integrates descriptive and prescriptive lines of research about randomness.

Normative, descriptive and prescriptive lines of research often study the same substantive content (e.g., CIs, correlation, randomness). However, the three lines of research have different goals, and are usually carried out by different scholars and published in different types of journals.

The Goals. Normative research aims to progress statistical theory, descriptive research aims to understand informal statistical reasoning, and prescriptive research aims to develop and evaluate improved strategies for teaching and practicing statistics.

The Scholars. Who are the people involved in this immense activity? Statisticians and mathematicians develop the science of statistics and so are most often responsible for the normative perspective. Cognitive psychologists contribute descriptive knowledge by studying how people reason statistically, interpret statistical concepts, make sense of statistical data; they describe people's correct or incorrect intuitions. Psychologists and educators in general, and teachers of statistics in particular, are often involved in studies aimed at improving statistical reasoning by suggesting new tools and methods of instruction (usually suggested by educators) or de-biasing techniques to overcome misconceptions (usually recommended by psychologists).

The Journals. Normative issues are mostly published in statistical journals, or in journals that focus on statistics and research methods in a particular discipline.

Statisticians publish in statistical journals (e.g., *Statistical Science*), while psychologists who are interested in normative issues often publish in the specialized psychological journals (e.g., *Psychological Methods*). Descriptive research on statistical reasoning often appears in psychology journals (e.g., *Journal of Behavioral Decision Making*, *Cognitive Psychology*), while prescriptive research is mostly seen in specialist statistics education journals (e.g., *The Journal of Statistics Education*, *Statistics Education Research Journal*, *Teaching Statistics*). These journals not only publish prescriptive research, but also to some extent have an agenda of integrating descriptive research, which is a laudable goal.

Regardless of the debate over when the modern era of statistical theory began, it is obvious that the normative tradition has a much longer history than either the descriptive or prescriptive traditions. Descriptive research dates back at least to the 1960s (e.g., Rosenthal & Gaito, 1963; Peterson & Beach, 1967). Prescriptive research on statistical practice and education is more recent, with formal societies and dedicated journals dating from around the mid 1980s.

We have demonstrated how often research in each of the three facets depends on the others and influences them. The organizational and sociological barriers between the three lines of research need to be removed, if EBP is to be achieved.

10. CONCLUSIONS

We have proposed the term statistical cognition for an integrative field that incorporates three lines of research. Interaction between the normative and prescriptive facets may seem relatively straightforward, so we focussed attention on the mutual contributions of the normative and descriptive facets, and the descriptive and prescriptive facets.

We discussed how normative models have been used as standards to which intuitive statistical reasoning (identified by descriptive research) is compared, with mismatches of varying extents emerging. The normative thus serves as a theoretical framework for describing how people *should* perform statistical tasks. We discussed NHST, CIs and Bayesian Hypothesis Testing as examples of how descriptive research on people's perception of statistical concepts can affect the normative status of models; this illustrates the contribution of descriptive research to the normative facet of statistical cognition.

Descriptive research on statistical reasoning aims to describe cognitive processes and misconceptions, and to detect developmental barriers to statistical reasoning. It can thus guide prescriptive investigations designed to identify the most efficient statistical training program. Conversely, prescriptive research on the effectiveness of the training program and teaching strategies, and the cognitive changes they elicit, provides empirical tests of descriptive models of people's statistical reasoning, thus enhancing or weakening their validity.

We used correlation as an example of how a statistical concept can be studied from all three perspectives of statistical cognition—its normative status, how people interpret it, and how it should be presented and explained—in order to improve statistics education and advising. We used CIs as an example of how research in all three facets might contribute to improving statistical practice. Finally, we identified barriers that we believe have hampered the interactions and synergies that are needed for EBP.

EBP is the key to better statistical practice and statistics education. It offers a number of advantages that should motivate its widespread adoption. Successful EBP

- ensures that consumers (in the fields we are discussing: researchers and students) get the best a discipline can offer;

- improves the efficiency of use of scarce resources, notably the time of teachers and other professionals;
- draws out practical implications from existing research for teachers and practitioners;
- guides the planning of future research; and
- encourages future research to generate practical implications.

A fear about EBP is that it might lead to a mechanistic, one-size-fits-all approach that marginalises the expertise and judgment of the teacher or statistician. However, recall the definition we quoted of medical EBP: "the integration of best research evidence with clinical expertise and patient values." We endorse this approach to EBP, which makes explicit the need for relevant expertise—of the teacher, statistician, or researcher—to ensure that lessons from the research evidence are applied appropriately, for maximum effect in a particular situation. Many factors influence decision making in statistical teaching and statistical consulting: statistical theory, ideology, values, clients, and personality factors. Even so, EBP can flourish.

In medicine, EBP has been primarily concerned with encouraging practitioners to make more use of research evidence that is already available. By contrast, in education greater emphasis has been placed on the absence of good quality research that can support EBP (Hargreaves, 1996). Davies (1999) emphasized that EBP in education should both draw on evidence from existing world-wide research and literature on education, and also encourage and guide further educational research.

In statistics education, there is already considerable descriptive and prescriptive research, and some integration of these two—for example in *SERJ*. However, there are also many gaps in the evidence needed to guide and justify EBP, and great scope for improved integration. It is not surprising that new research fields develop and specialize, building their own institutions, journals, and cultures. However, to adopt EBP in a thorough way requires reintegration, which both facilitates mutual contributions, and helps identify serious gaps in current knowledge. Reintegration is essential—and we believe the umbrella of *statistical cognition* can be very helpful—for building a cohesive and complete evidence base.

Why should labelling this integration and introducing the umbrella term help? It may, for example, help remind researchers, as they embark on a prescriptive or normative research program, to think also of relevant descriptive research that impacts on their goals—and vice versa. The term 'statistical cognition' in particular highlights the importance of a cognitive evidence base as well as a statistical and a pedagogical one.

For statistical practice, and especially for the reformed practice needed in psychology and other disciplines still over-reliant on NHST, the descriptive evidence base is very sparse, and very little prescriptive research has been conducted. There is enormous scope for a statistical cognition perspective to encourage and guide research, and to build the integrated evidence base needed for improved statistical practice and statistics education. The organizational and sociological factors responsible for the barriers between the three facets should now be exploited to overcome them. We look forward to further discussion of statistical cognition—and perhaps to the emergence of an international conference and a journal titled *Statistical Cognition*—and the potential we believe it has to energise and support the expansion of EBP in statistical practice and statistics education.

ACKNOWLEDGMENTS

This research was conducted while Ruth Beyth-Marom was a Visiting Fellow at the Institute of Advanced Studies, La Trobe University.

REFERENCES

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (2000). *Statistics with confidence* (2nd ed.). London: British Medical Journal.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management, 64*, 912-923.

Bar-Hillel, M. (1979). The role of sample size in sample evaluation. *Organizational Behavior and Human Performance, 24*, 245-257.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*, 211-233.

Bar-Hillel, M., & Neter, E. (1993). How alike is it versus how likely is it: A disjunction fallacy in probability judgments. *Journal of Personality and Social Psychology, 65*, 1119-1131.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*, 389-396.

Bell, D. E., Raiffa, H., & Tversky, A. (1988). *Decision making: Descriptive, normative, and prescriptive interactions*. New York: Cambridge University Press.

Ben-Zvi, D. & Garfield, J. (Eds.). (2004). *The challenge of developing statistical literacy, reasoning and thinking*. Dordrecht, The Netherlands: Kluwer.

Beyth-Marom, R. (1982). Perception of correlation reexamined. *Memory & Cognition, 10*, 511-519.

Beyth-Marom, R. (1990). Mis/understanding diagnosticity: Direction and magnitude of change. In K. Borcherding, O. L. Larichev & D. M. Mesick (Eds.), *Contemporary issues in decision making* (pp. 203-223). North Holland: Elsevier.

Beyth-Marom, R., & Fischhoff, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality and Social Psychology, 45*, 1185-1195.

Biggs, J., & Collis, K. (1991). Multimodal learning and the quality of intelligent behavior. In H. Rowe (Ed.), *Intelligence, Reconceptualization and Measurement* (pp. 57-76). Hillsdale, NJ: Erlbaum.

Brown, J. S., Collings, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*, 32-41

Cockcroft, W. H. (1982). *Mathematics counts: Report of the committee of inquiry into the teaching of mathematics in schools*. London: HMSO.

Cohen, J. (1994). The earth is round (*p*<.05). *American Psychologist, 49*, 997-1003.

Cohen, L. J. (1979). On the psychology of prediction: Whose is the fallacy? *Cognition, 7*, 385-407.

Cumming, G. (2007). Inference by eye: Pictures of confidence intervals and thinking about levels of confidence. *Teaching Statistics, 29*, 89-93.

Cumming, G., & Fidler, F. (in press). The new stats: Effect sizes and confidence intervals. In G. R. Hancock & R. O. Mueller (Eds.) *Quantitative methods in the social and behavioral sciences: A guide for researchers and reviewers*. Hillsdale, NJ: Erlbaum.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170-180.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*, 299-311.

Davies, H. T. O. (1999). What is evidence based education? *British Journal of Educational Studies, 47,* 108-121.

Donovan, M. S., Bransford, J. D., & Pellegrino, J. W. (Eds.). (1999). *How people learn: Bridging research and practice*. Washington, DC: National Academy Press.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.). *Judgment under uncertainty: Heuristics and biases* (pp. 249-267). New York: Cambridge University Press.

Edwards, A. L. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.). *Formal representation of human judgment* (pp. 17-52). New York: Wiley.

Engel Clough, E., & Wood-Robinson, C. (1985). How secondary students interpret instances of biological adaptation. *Journal of Biology Education, 19*, 125-130.

Erickson, G. L. (1980). Children's viewpoints of heat: A second look. *Science Education, 64*, 323-336.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning*, *9*, 83-96.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard. *Theory and Psychology, 5,* 75-98.

Falk, R., & Konold, C. (1994). Random means hard to digest. *Focus on Learning Problems in Mathematics, 16*, 2-12.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104*, 301-318.

Fidler, F., & Cumming, G. (2008). The new stats: Attitudes for the twenty-first century. In J. W. Osborne (Ed.). *Best practices in quantitative methods* (pp. 1-12). Thousand Oaks, CA: Sage.

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., et al. (2005). Toward improved statistical reporting in the *Journal of Consulting and Clinical Psychology*. *Journal of Consulting and Clinical Psychology, 73*, 136-143.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement, 61,* 181-210.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*, 239-260.

Fisher, K. (1985). A misconception in biology: Amino acids and translation. *Journal of Research in Science Teaching, 22*, 53-62.

Gal, I. (2002). Adults' statistical literacy: Meanings, components and responsibilities. *International Statistical Review, 70*, 1-25.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education, 10*(3).
[Online: http://www.amstat.org/publications/jse/v10n3/garfield.html]

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal of Research in Mathematics Education, 19*, 44-63.

Garfield, J., & Gal, I. (1999), Teaching and assessing statistical reasoning. In L. Stiff (Ed.), *Developing mathematical reasoning in grades K*-12 (pp. 207-219). Reston, VA: National Council Teachers of Mathematics.

Garfield, J., Hogg, B., Schau, C., & Whittinghill, D. (2002). First courses in statistical science: The status of educational reform efforts. *Journal of Statistics Education 10*(2).
[Online: www.amstat.org/publications/jse/v10n2/garfield.html]

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond heuristics and biases. *European Review of Social Psychology, 2*, 83-115.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684-704.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgments.* Cambridge, UK: Cambridge University Press.

Glaser, R., & Bassok, M. (1989). Learning theory and the study of instruction. *Annual Review of Psychology, 40*, 631-666.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41-58). San Diego, CA: Academic Press.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*, 1-20.

Hargreaves, C. (1996). *Teaching as a research based profession: Possibilities and prospects*. London: Teacher Training Agency.

Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin, 118*, 248-271.

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology and its future prospects. *Educational and Psychological Measurement, 60,* 661-681.

Huff, D. (1973). *How to lie with statistics*. Harmondsworth: Penguin.

Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.

Jenkins, H., & Ward, W. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs, 19*, 1-17.

Jones, G. A., Langrall, C. W, Thornton, C. A., & Mogill, A. T. (1997). A framework for assessing and nurturing young children's thinking in probability. *Educational Studies in Mathematics, 32*, 101-125.

Jones, G. A., Langrall, C. W., Thornton, C. A., & Mogill, A. T. (1999). Students' probabilistic thinking in instruction. *Journal for Research in Mathematics Education, 30*, 487-519.

Kafoussi, S. (2004). Can kindergarten children be successfully involved in probabilistic tasks? *Statistics Education Research Journal, 3*(1), 29-39.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(1)_kafoussi.pdf]

Kahneman, D. (2003). A perspective on intuitive judgment and choice: Maps of bounded rationality. *American Psychologist, 58*, 697-720.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.

Kalinowski, P., Fidler, F., & Cumming, G. (2008). *Overcoming the inverse probability fallacy: A comparison of two teaching interventions*. Manuscript in preparation.

Kitson, A., Harvey, G., & McCormack, B. (1998). Enabling the implementation of evidence based practice: A conceptual framework. *Quality in Health Care, 17,*149-158.

Kline, R. B. (2004). *Beyond significance testing. Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Lawrenz, F. (1986). Misconceptions of physical science concepts among elementary school teachers. *School Science and Mathematics, 86*, 654-660.

Leron, U., & Hazzan, O. (2006). The rationality debate: Application of cognitive psychology to mathematics education. *Educational Studies in Mathematics*, *62,* 105-126.

Meehl, P. (1954). *Clinical versus statistical prediction.* Minneapolis, MN: University of Minnesota Press.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806-834.

Mehlinger, H. D. (1995). *School reform in the information age.* Bloomington, IN: Indiana University Press.

Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning, 4*, 23-63.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Oxford English Dictionary. (2007). Retrieved on October 15, 2007 from http://www.oed.com

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*, 29-46.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. London: Routledge & Kegan Paul.

Pinker, S. (2002). *The blank slate: The modern denial of human nature*. New York: Viking Penguin.

Pollatsek, A., Well, A. D., Konold, C., Hardiman, P., & Cobb, G. (1987). Understanding conditional probabilities. *Organizational Behavior and Human Decision Processes, 40*, 255-269.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology, 55,* 33–38.

Rossi, J. S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 175-197). Hillsdale, NJ: Lawrence Erlbaum.

Schmidt, F. L. & Hunter, J. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-63). Hillsdale, NJ: Lawrence Erlbaum.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications.* London: LEA publishers.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 107*, 309-316.

Shaklee, H., & Tucker, D. (1980). A rule analysis of judgment of covariation between events. *Memory & Cognition, 8*, 459-467.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3-22.

Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology, 4*, 165-173.

Smith, E. L., & Anderson, C. W. (1984). Plants as producers: A case study of elementary science teaching. *Journal of Research in Science Teaching, 21*, 685-698.

Smith, G. (1998). Learning statistics by doing statistics. *Journal of Statistics Education, 6*, 1-12.
[Online: http://www.amstat.org/publications/jse/v6n3/smith.html]

Stanovich, K.E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23*, 645-726.

The American Heritage Dictionary the English Language (4th ed.). (2000). Boston: Houghton Mufflin.

Trinder, L., & Reynolds, S. (Eds.) (2000). *Evidence-based practice: A critical appraisal.* London: Blackwell Science.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science 185*, 1124-31.

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., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293-315.

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association, 88*, 1-8.

Ward, W., & Jenkins, H. (1965). The display of information and judgment of contingency. *Canadian Journal of Psychology, 19*, 231-241.

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3-46.
[Online: www.stat.auckland.ac.nz/~iase/serj/SERJ2(2)_Watson_Callingham.pdf]

Webster's Online. (2007). Retrieved Oct. 15, 2007 from www.websters-dictionary-online.org

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

RUTH BEYTH-MAROM
Department of Education and Psychology
The Open University of Israel
108 Ravutski St. Raanana, 43107 Israel

A FRAMEWORK TO SUPPORT RESEARCH ON INFORMAL INFERENTIAL REASONING

ANDREW ZIEFFLER
University of Minnesota
zief0002@umn.edu

JOAN GARFIELD
University of Minnesota
jbg@umn.edu

ROBERT DELMAS
University of Minnesota
delma001@umn.edu

CHRIS READING
University of New England
creading@une.edu.au

ABSTRACT

Informal inferential reasoning is a relatively recent concept in the research literature. Several research studies have defined this type of cognitive process in slightly different ways. In this paper, a working definition of informal inferential reasoning based on an analysis of the key aspects of statistical inference, and on research from educational psychology, science education, and mathematics education is presented. Based on the literature reviewed and the working definition, suggestions are made for the types of tasks that can be used to study the nature and development of informal inferential reasoning. Suggestions for future research are offered along with implications for teaching.

Keywords: *Statistics education research; Inference; Informal reasoning; Introductory statistics course; Topic sequencing*

1. INTRODUCTION

> Statistics is concerned with the gathering, organization, and analysis of data and with inferences from data to the underlying reality. (Moore, 1990, p. 127)

Drawing inferences from data is part of everyday life and critically reviewing results of statistical inferences from research studies is an important goal for most students who enroll in an introductory statistics course. Formal methods of statistical inference lead to drawing conclusions about populations or processes based on sample data. David Moore (2004) describes statistical inference as moving beyond the data in hand to draw conclusions about some wider universe, taking into account that variation is everywhere and the conclusions are therefore uncertain. Garfield and Ben-Zvi (2008) define statistical inference further by differentiating two important themes in statistical inference,

Statistics Education Research Journal, 7(2), 40-58, http://www.stat.auckland.ac.nz/serj

parameter estimation and hypothesis testing, and two kinds of inference questions, generalization (from samples) and comparison and determination of cause (from randomized comparative experiments). In general terms, the first theme is concerned with generalizing from a small sample to a larger population, whereas the second involves determining whether a pattern in the data can be attributed to a real effect.

For several decades, psychologists and education researchers have studied and documented the difficulties people have making inferences about uncertain outcomes (see Kahneman, Slovic, & Tversky, 1982). Falk and Greenbaum (1995) and others (e.g., Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007) have described and classified difficulties students have in understanding and interpreting tests of significance and p-values and related concepts in statistical inference. Researchers have pointed to various reasons for these difficulties including: the logic of statistical inference (e.g., Cohen, 1994; Nickerson, 2000; Thompson, Saldanha, & Liu, 2007), students' intolerance for ambiguity (Carver, 2006), and students' inability to recognize the underlying structure of a problem (e.g., Quilici & Mayer, 2002). Other research has suggested that students' incomplete understanding of foundational concepts such as distribution (e.g., Bakker & Gravemeijer, 2004), variation (e.g., Cobb, McClain, & Gravemeijer, 2003), sampling (e.g., Saldanha & Thompson, 2002, 2006; Watson, 2004), and sampling distributions (e.g., delMas, Garfield, & Chance, 1999; Lipson, 2003) may also play a role in these difficulties.

Given the importance of understanding and reasoning about statistical inference, and the consistent difficulties students have with this type of reasoning, there have been attempts to expose students to situations that allow them to use informal methods of making statistical inferences (e.g., comparing two groups based on boxplots of sample data). Several papers have been presented and published in the past few years that describe 'informal statistical inference' and 'informal inferential reasoning' (e.g., Pfannkuch, 2005). However, it is not yet clear exactly what these two terms mean. Therefore, it is the intent of this paper to analyze the meaning of Informal Inferential Reasoning (IIR) by reviewing the literature related to this topic, and to provide both a working definition as well as a framework for designing tasks that can be used to study students' reasoning about statistical inference.

Cognitive frameworks have been useful in studying and describing students' statistical reasoning (see Jones, Langrall, Mooney, & Thornton, 2005). These models offer benchmarks for assessing students' reasoning and are useful for informing the development of assessment tasks, guiding teachers' instructional decision-making, and developing tasks to use in research programs. The main focus of this paper is to propose a preliminary framework that, although not a developmental model, can be used to identify and develop tasks that can be used to study IIR. The two main questions addressed in the paper are

1. What are the components of a framework needed to support research on informal inferential reasoning?
2. What types of tasks are suggested by this framework for the study of informal inferential reasoning and its development?

2. WHAT ARE THE COMPONENTS OF A FRAMEWORK NEEDED TO SUPPORT RESEARCH ON INFORMAL INFERENTIAL REASONING?

The previous section described the nature of statistical inference and the way concepts and procedures involved in statistical inference are often introduced in introductory statistics courses. In an attempt to understand the component parts of informal inferential reasoning (IIR), we look to the literature to identify some foundational areas of research. Because IIR uses the word "informal" it seemed useful to explore research in two possibly related areas of research in psychology and education: studies of informal knowledge and studies of informal reasoning. The research in both of these areas provides a foundation for understanding and defining IIR. This section begins with a brief review of the research in the areas of informal knowledge and then informal reasoning. This is followed by a review of the use of the terms "informal inference" and "informal inferential reasoning" by statistics educators and statistics education researchers in recent papers. The section concludes with a working definition of IIR based on the literature reviewed in these three areas.

2.1. INFORMAL KNOWLEDGE

There is much research on the nature of informal knowledge, particularly in the field of mathematics education. Informal knowledge is viewed as either a type of *everyday real world knowledge* that students bring to their classes based on out-of-school experiences, or a less formalized knowledge of topics resulting from prior formal instruction. Informal knowledge can be viewed as the integration of both of these and it is in this sense that the term is used in this paper. This view suggests that it is important to study and consider the role of informal knowledge in the formal study of a particular topic and is in line with constructivist views of learning, namely that informal knowledge is a starting point for the development of formal understanding.

Informal knowledge is also discussed in the literature on how experts use their informal knowledge in reasoning while solving problems, and in studies that compare the reasoning of experts and novices. Smith, diSessa and Rochelle (1993/1994) found that experts appear to differ from novices primarily in the extent of their experience with problems in a particular context. Although experts seemed to see a coherent picture of a problem they were solving, novices were more likely to approach each problem in a similar group from a different starting point, rather than seeing how the problems were related. Experts also tended to draw on their more extensive experiences and knowledge to use problem-solving strategies that related to the underlying structures of the problems more often than novices. For a more detailed synthesis of the research on experts versus novices see Bransford, Brown, and Cocking (2000).

When informal knowledge is incorrect, it is often regarded as a misconception (see Confrey, 1990). Smith et al. (1993/1994) argued that it is not beneficial to view students' "misconceptions" as wrong, because there are aspects of students' informal knowledge that are similar to experts' knowledge, but have been used incorrectly. Novice reasoning tends to have the same basic structure as expert reasoning, but novice reasoning often appears more concrete and less abstract due to novices' more limited experience. Instead of focusing only on the development of formal knowledge, Smith et al. suggest that instructors carefully design lessons that build and develop students' informal knowledge in order to lead them toward formal understanding of a particular topic.

Along the same lines, Gravemeijer and Doorman (1999) argue that it is important to have students build on their informal knowledge to reinvent formal concepts and representations and at the same time expand their common sense understanding of real world phenomena. This approach acknowledges, rather than discounts (e.g., by labeling erroneous use of knowledge as misconceptions), the informal knowledge that students bring to the classroom.

An important question emerges: How can students' informal knowledge best be utilized in formal instruction? Some researchers point to the role of interactive activities where students work and discuss together what they are learning. It has been suggested and demonstrated that social interaction that requires negotiation of meaning, under the direction of shared social norms for communication helps support the transformation of informal knowledge to culturally shared formal understanding (Cobb & McClain, 2004; Cobb, Yackel, & Wood, 1992; Mack, 1995).

Another way of developing students' informal knowledge is to specifically evolve this type of knowledge through activities that motivate and "set the stage" for formal instruction at a later time (see for example, Papert & Harel, 1991). Schwarz, Sears, and Chang (2007) have found positive results in their attempts to explicitly develop and utilize students' prior knowledge as they learn specific statistical concepts. Garfield, delMas, and Chance (2007) have also found some success in developing college students' formal ideas of variability from informal ideas.

In summary, the literature reviewed suggests that

1. Informal knowledge can consist of different types of understanding that students bring to a new learning task, and may combine knowledge based on real world experience with knowledge gained from previous instruction (Gravemeijer & Doorman, 1999; Smith et al., 1993/1994).
2. Informal knowledge may be an important starting point on which to build formal knowledge, and should be considered in designing curricula (Gravemeijer & Doorman, 1999; Smith et al., 1993/1994).
3. Instruction may be designed to help students construct specific types of informal knowledge that is needed for eventual instruction involving formal knowledge of a particular concept (Garfield, delMas, & Chance, 2007; Schwarz, Sears, & Chang, 2007).
4. Activity-based learning that requires social interaction and the negotiation of meaning can facilitate the development of informal knowledge (Cobb & McClain, 2004; Cobb, Yackel, & Wood, 1992; Mack, 1995).

This review of the nature of informal knowledge suggests that developing students' informal knowledge related to statistical inference may ease their transition to understanding formal ideas of inference.

2.2. INFORMAL REASONING

Informal reasoning (sometimes referred to as informal logic) has been defined by cognitive psychologists as the type of reasoning that occurs in non-deductive situations, such as decision making, that is employed in everyday life (Voss, Perkins, & Segal, 1991). Perkins, Farady, and Bushey (1991) characterize informal reasoning as "a process of *situation modeling*" in which a person builds a model of the situation in question by "articulating the dimensions and factors involved … and invok[ing] a variety of common sense, causal, and intentional principles both to construct and to weigh the plausibility of alternative scenarios" (p. 85).

Formally, there is very little agreement in the literature as to what is meant by 'informal reasoning'. This may be due to the reliance of informal reasoning on context or subject matter (Perkins, 1985b; Walton, 1989). There are, however, two commonalities across a majority of the papers reviewed. First, informal reasoning is most often viewed through the lens of argumentation theory (e.g., Kuhn, 1991; Means & Voss, 1996; Sadler, 2004; Sadler & Zeidler, 2004). Secondly, informal reasoning is often contrasted to formal reasoning or logic (e.g., Evans, Newstead, & Byrne, 1993; Miller-Jones, 1991; Pfannkuch, 2006; Schoenfeld, 1991).

Researchers tend to study informal reasoning through dialogical argumentation–the expression or means by which researchers gain access to informal reasoning (e.g., Driver, Newton, & Osborne, 2000; van Eemeren et al., 1996). The assessment of informal reasoning through argumentation, which draws heavily from Toulmin's (1958) model of argumentation, has led to many findings. For example, Perkins et al. (1991) have found through a series of studies that informal reasoning, despite claims to the contrary, "is marred by incompleteness and bias is the norm rather than the exception" (p. 90). Sadler and Zeidler (2004) caution that "while it is valid to assert that strong argumentation reveals strong informal reasoning, the opposite claim, weak argumentation denotes weak informal reasoning, is not necessarily the case … naïve arguments might be the result of either insufficient informal reasoning or poorly articulated, but proficient informal reasoning" (p. 73).

In summary, the literature reviewed suggests that

1. The quality of informal reasoning does not necessarily improve with increased content knowledge (e.g., Kuhn, 1991; Means & Voss, 1996; Perkins, 1985b; Perkins et al., 1991);
2. Informal reasoning is unlikely to improve with maturation, education, or life experience (e.g., Klahr, Fay, & Dunbar, 1993; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995; Perkins, 1985b; Perkins et al., 1991; Schauble, 1990, 1996; Schauble & Glaser, 1990; Voss, Blais, Means, Greene, & Ahwesh, 1986);
3. Motivation, or interest in the problem context, has little impact on informal reasoning quality (e.g., Perkins, 1989; Perkins et al., 1991);
4. General intelligence influences people's informal reasoning, but people selectively use that intelligence to build their own case rather than to explore an issue more fully (e.g., Perkins, 1985a; Perkins, 1989; Perkins et al., 1991);
5. Informal reasoning is a matter of "know-how" and can be improved through instruction (e.g., Nickerson, Perkins, & Smith, 1985; Perkins, Bushey, & Farady, 1986; Perkins et al., 1991; Schoenfeld, 1982; Schoenfeld & Herrmann, 1982).

Informal reasoning seems to be an important part of IIR because of the role of evidence and argumentation in making statistical predictions and decisions.

2.3. DEFINING INFORMAL INFERENTIAL REASONING

As mentioned earlier, IIR is a relatively recent concept in the research literature and various definitions have been presented. Rubin, Hammerman, and Konold (2006) define IIR as reasoning that involves the related ideas of properties of aggregates (e.g., signal and noise, and types of variability), sample size, and control for bias. Pfannkuch (2006) defines IIR as the ability to interconnect ideas of distribution, sampling, and center, within an empirical reasoning cycle (Wild & Pfannkuch, 1999). Bakker, Derry, and Konold (2006) suggest a theoretical framework of inference that broadens the meaning of statistical inference to allow more informal ways of reasoning and to include human judgment based on contextual knowledge. One statistician has described informal inference as "going beyond the data at hand" and "seeking to eliminate or quantify chance as an explanation for the observed data" through a reasoned argument that employs no formal method, technique, or calculation (Rossman, 2007). Ben-Zvi (2006) compares inferential reasoning to argumentation, and emphasizes the need for this type of reasoning to include data-based evidence.

These different definitions of IIR share many things in common. In an attempt to combine these perspectives, we present a working definition of informal inferential reasoning as *the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples.* In addition, building on their own experiences and understanding from teaching formal ideas and methods of statistical inference to college students, the authors see informal inferential reasoning as a process that includes

- Reasoning about possible characteristics of a population (e.g., shape, center) based on a sample of data;
- Reasoning about possible differences between two populations based on observed differences between two samples of data (i.e., are differences due to an effect as opposed to just due to chance?); and
- Reasoning about whether or not a particular sample of data (and summary statistic) is likely (or surprising) given a particular expectation or claim.

This is in contrast to formal statistical inferential reasoning, which may include significance tests and/or confidence intervals. For example, the type of formal reasoning about a one-sample test of significance that we hope to help students eventually develop requires an understanding of the interconnections between

- An underlying theory or hypothesis that is to be tested;
- A sample of data that can be examined; and
- A distribution of a statistic for all possible samples under the assumption that the theory or hypothesis is true.

This integration involves comparing the observed sample statistic to the distribution of statistics for all possible samples to see how unlikely the occurrence is (i.e., how far out in either of the tails it falls). The farther out in one of the tails, the less plausible it is that the observed results are due to chance and, therefore, the more convincing that there is a true difference or effect. This formal reasoning also includes the understanding of a p-value as an indicator of how likely or surprising a sample result, or a result more extreme, is under a certain hypothesis, and the action of rejecting this hypothesis if the p-value is small enough.

In summary, the IIR Framework has the following three components:

1. Making judgments, claims, or predictions about populations based on samples, but not using formal statistical procedures and methods (e.g., p-value, *t* tests);
2. Drawing on, utilizing, and integrating prior knowledge (e.g., formal knowledge about foundational concepts, such as distribution or average; informal knowledge about inference such as recognition that a sample may be surprising given a particular claim; use of statistical language), to the extent that this knowledge is available; and
3. Articulating evidence-based arguments for judgments, claims, or predictions about populations based on samples.

Note that this definition refers to IIR as a process for making inferences that does not utilize the formal methods of statistical inference described earlier and that may or may not include use of formal statistical concepts or language.

2.4. WHY STUDY INFORMAL INFERENTIAL REASONING?

Given the importance of statistical inferential reasoning, and given the difficulties most people have with this type of reasoning, a better pedagogical approach to this topic is needed. One possible cause for students' difficulty with formal statistical inferential reasoning is that they lack both experience with stochastic events that form the underpinnings of statistical inference (Pfannkuch, 2005), and experience of reasoning about these events. Statistics educators and statistics education researchers have recently been exploring the idea that if students begin to develop the informal ideas of inference (as defined above) early in a course or curriculum, they may be better able to learn and reason about formal methods of statistical inference. For example, early on in a course, students could engage in discussions of when, compared to some agreed upon expectation, a sample is surprising. This discussion could be revisited during different activities that introduce different samples and distributions. Over time, this may develop the prior statistical knowledge needed to understand the idea of a p-value when it is introduced later in a course.

There is also a belief that, if students become familiar with reasoning about inference in an informal manner, such as making speculations about what might be true in a population or populations, based on samples of data, that the method of doing this formally may be more accessible. Finally, because statistical inference integrates many important ideas in statistics—such as data representation, measures of center and variation, the normal distribution, and sampling—introducing informal inference early and revisiting the topic throughout a single course or curriculum across grades could provide students with multiple opportunities to build the conceptual framework needed to support inferential reasoning. These suggestions for ways in which students reason informally about statistical inference, as well as possible methods for developing that reasoning, are currently untested conjectures that need to be studied.

Now that a working definition of IIR and a rationale for studying IIR have been provided, many questions emerge. For example, how can researchers investigate and describe the nature of IIR in students? Or, what are ways to challenge students to reveal their informal inferential reasoning? The three components of the IIR framework can be used to help answer these questions because they support the development of tasks to examine students' intuitive IIR as well as their developing reasoning.

3. WHAT TYPES OF TASKS CAN BE USED TO STUDY INFORMAL INFERENTIAL REASONING AND ITS DEVELOPMENT?

The research literature contains examples of two different approaches that researchers have used to study IIR. One approach focuses on the nature of this reasoning or naïve methods of reasoning about inference given problems and statistical information. An objective of this type of study is often to examine how students reason about or make inferences given a particular problem without having encountered formal methods of statistical inference via instruction. A second approach is the examination of the development of IIR as students experience curricula (e.g., a course or unit of instruction) designed to build reasoning. The objective of this type of study is often to see how the nature of students' inferential reasoning changes as they are provided with resources, tools, and curriculum. Both of these approaches need well-designed tasks that allow researchers to capture and evaluate students' IIR.

Reading (2007) suggested that tasks used in a study of students' informal inferential reasoning would not only need to examine how students integrate the components of the IIR framework listed in Section 2.3, but also capture ideas of statistical inference such as generalizing to an appropriate population beyond a collected sample, basing inferences on evidence, choosing between competing models (i.e., hypotheses), expressing a degree of uncertainty in making an inference, and making connections between the results and problem context. Furthermore, the research literature on informal reasoning and informal knowledge would suggest that tasks should be designed to elicit multiple arguments from students, as well as separate novice reasoning from expert reasoning.

The framework provided in the working definition in this paper suggests the design of tasks that challenge students to

1. Make judgments, claims, or predictions about a population based on samples, but not using formal statistical procedures and methods (e.g., p-value, *t* tests);

2. Draw on, utilize, and integrate prior knowledge (formal and informal) to the extent that this knowledge is available; and
3. Articulate evidence-based arguments for judgments, claims, and predictions about populations based on samples.

Three general types of task have been used in research studies that meet these criteria. They may be categorized as tasks that ask students to

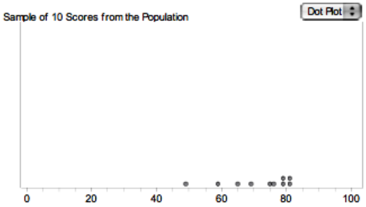
1. Estimate and draw a graph of a population based on a sample;
2. Compare two or more samples of data to infer whether there is a real difference between the populations from which they were sampled, and
3. Judge which of two competing models or statements is more likely to be true.

Examples of tasks for each of three categories are provided in the following sections.

3.1. ESTIMATE AND DRAW A GRAPH OF A POPULATION BASED ON A SAMPLE

Very few examples appear in the literature where students have been asked to speculate about graphical characteristics of a population based on a sample of data. Bakker (2004) referred to this type of prediction as “growing a sample” and used the following task in a teaching experiment with eighth grade students in the Netherlands. Students were asked to predict a graph of weights for a class of 27 eighth grade students and then graphs for three classes together, which had a total of 67 students, based on small random samples of student weights. After being shown the computer-simulated data sets for one class of 27 students and all three classes together, they were asked to describe the differences between their two graphs and then to compare these to the real graphs of weight data. In the last part of the activity, students were asked to create graphs for the population of all students in their city that were no longer sets of points but were continuous distributions of data. This multistage activity ended in an IIR activity that had students make a conjecture about an unknown population. Based on Bakker’s activity, the following task (see Figure 1) was created and used by Zieffler, delMas, Garfield, and Gould (2007) to reveal students’ IIR in an introductory college statistics course.

Imagine the test scores for a group of college students in a very large lecture class on psychology (*n*=1000 students). The test scores for a random sample of ten students from this class are shown in the dot plot [below].



- Now, consider a random sample of 25 students drawn from the same class. Try to imagine what THAT graph might look like. Use the graphing area to sketch a dot plot of the 25 scores that you might expect to see for a random sample of 25 students. Explain your reasoning.
- Next, think about the entire class of 1000 students that these students are sampled from. What would you expect the distribution for the entire population of all 1000 students test scores to look like? Draw an outline of the distribution and explain your reasoning.

Figure 1. Predicting characteristics of a population from a sample task (Zieffler et al., 2007)

3.2. COMPARE TWO OR MORE SAMPLES OF DATA TO INFER WHETHER THERE IS A REAL DIFFERENCE BETWEEN THE POPULATIONS

There are many examples in the literature of tasks that ask students to compare two or more groups of data, although not all of them ask the students to reason beyond the samples to the populations from which they have been selected. For example, Watson and Moritz (1999) used tasks in which students in grades 3 through to 9 had to compare two data sets to help them begin to make inferences about group differences. Although these tasks did not look beyond the data sets to larger populations, they set the stage for such inferences, providing a foundation for statistical inference. One such task is in Figure 2.

Two schools are comparing some classes to see which is better at quick recall of 9 maths facts. In each part of this question you will be asked to compare different classes. First consider two classes, the Blue class and the Red class. The scores for the two classes are shown on the two charts below. Each box is one person's test, and the number inside is their score. In the *Blue* class, 4 people scored 2 correct and 2 people scored 3. In the *Red* class, 3 people scored 6 correct and 3 people scored 7.

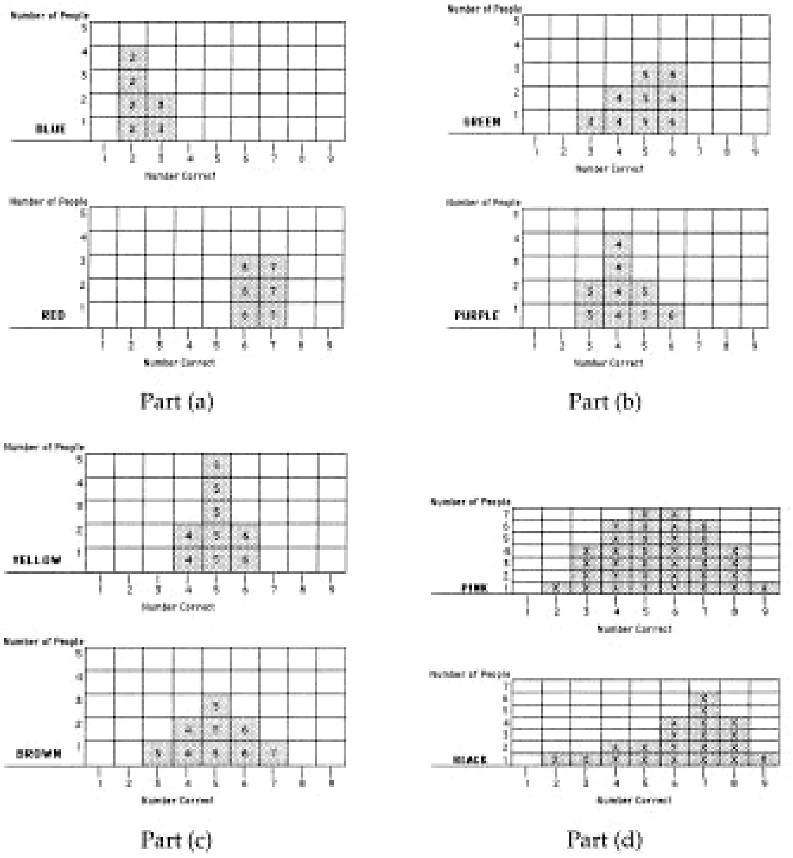


Figure 2. Comparing groups task (Watson & Moritz, 1999)

In contrast, Pfannkuch (2005, 2006) used a task that asked students to compare sets of data for daily temperatures for two different cities in New Zealand, and challenged the students to make some informal inferences beyond the sample data. According to Pfannkuch (2005), "students were required to pose a question (e.g., Which city has the higher maximum temperatures in summer?), analyze the data, draw a conclusion, justify the conclusion with three supporting statements, and evaluate the statistical process" (p. 272). All students constructed a pair of boxplots to make the comparison (see Figure 3).

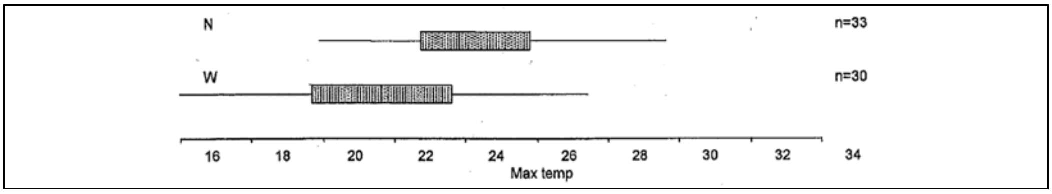
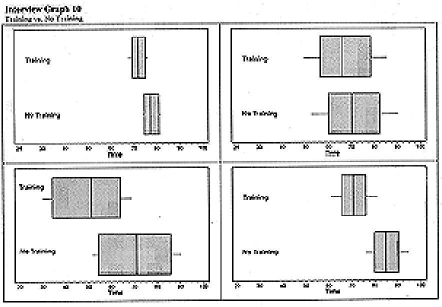


Figure 3. Boxplots of temperature data from comparing groups task (Pfannkuch, 2005)

Suppose that there is a special summer camp for track athletes. There is one group of 100 athletes that run a particular race, and they are all pretty similar in their height, weight, and strength. They are randomly assigned to one of two groups. One group gets an additional weight-training program. The other group gets the regular training program without weights. All the students from both groups run the race and their times are recorded, so that the data could be used to compare the effectiveness of the two training programs.

- Describe what you would expect to see in a comparison of two graphs if the difference between the two groups of athletes is really not due to the training program.
- Describe what you would expect to see in a comparison of two graphs if the difference between the two groups of athletes is really due to the training program.

Presented below are some possible graphs that show boxplots for different scenarios, where the running times are compared for the students in the two different training programs (one with weight training and one with no weight training). Examine each pair of graphs and think about whether or not the sample data would lead you to believe that the difference in running times is caused by these two different training programs. (Assume that everything else was the same for the students and this was a true, well-designed experiment.)



- Which set of boxplots show the MOST convincing evidence that the weight-training program was more effective in DECREASING athlete's running times? Explain.
- Which set of boxplots shows the LEAST convincing evidence that the weight-training program was more effective? Explain.
- Rank the four pairs of graphs on how convincing they are in making an argument that the weight-training program was more effective in decreasing athletes' times (from the least convincing to the most convincing evidence). Explain your reasoning.
- For the pair of graphs that provide the most convincing evidence, would you be willing to generalize the effects of the training programs to all similar athletes on track teams, based on these samples? Why or why not?

Figure 4. Comparing groups task (Zieffler et al., 2007)

A third type of task has students compare multiple pairs of sample data and has the students judge which pair provides the most compelling evidence to support a true difference in population means. For example, Zieffler et al. (2007) used a task (Figure 4) to have students make conjectures about whether or not a special athletic training program was effective, based on two samples of data. This task required students to look beyond the samples of data to compare two populations, and to choose which pair of samples gave the most compelling evidence to support a claim that there was a real difference.

3.3. JUDGE WHICH OF TWO COMPETING MODELS OR STATEMENTS IS MORE LIKELY TO BE TRUE

A review of the literature found three different styles of task that have been used to challenge students to choose between two competing models or claims, based on sample data. One style uses data generated by a probability device, and uses proportions or percentages to summarize the sample data. For example, Stohl and Tarr (2002) and Tarr, Stohl Lee, and Rider (2006) used the *Schoolopoly* problem (Figure 5) that asked sixth-grade students to judge whether a die is fair or not based on observed tosses via a computer simulation. Students were asked to provide what they "consider 'compelling evidence' in formulating and evaluating arguments based on data" (Tarr, Stohl Lee, & Rider, 2006, p. 1).

Schoolopoly: Is the die fair or biased?

Background

Suppose your school is planning to create a board game modeled on the classic game of *Monopoly*. The game is to be called *Schoolopoly* and, like *Monopoly*, will be played with dice. Because many copies of the game expect to be sold, companies are competing for the contract to supply dice for *Schoolopoly*. Some companies have been accused of making poor quality dice and these are to be avoided since players must believe the dice they are using are actually "fair." Each company has provided dice for analysis and you will be assigned one company to investigate:

Luckytown Dice Company
Dice R' Us
High Rollers, Inc.
Dice, Dice, Baby!
Pips and Dots
Slice 'n' Dice

Your Assignment

Working with a partner, investigate whether the dice sent to you by the company are *fair* or *biased*. That is, collect data to infer whether all six outcomes are equally likely and answer the following questions:

1. Do you believe the dice you tested are fair or biased? Would you recommend that dice be purchased from the company you investigated?
2. What *compelling evidence* do you have that the dice you tested are fair or unfair?
3. Use your data to estimate the probability of each outcome, 1-6, of the dice you tested.

Collect data about the dice supplied to you. Note that each single trial represents the outcome of one roll of a "new" virtual die provided by the company.

Copy any graphs and screen shots you want to use as evidence and print them for your poster. Give a presentation pointing out the highlights of your group's poster.

Figure 5. Competing models task: Schoolopoly (Stohl & Tarr, 2002)

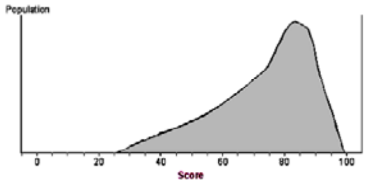
A second style of competing models task was used by Rubin, Hammerman, and Konold (2006). They had teachers determine whether a particular change in process either occurred or did not occur during a specified interval of time. This task is described in Figure 6.

The Mus-Brush Company produces mushroom brushes, using a large machine whose output is on average 215 brushes every two minutes *if it is working normally*. If the electricity to the machine is interrupted, even for a brief time, it will slow down such that the output of the machine will be 10% lower on average. The Mus-Brush Company was robbed last night; in forcing the door open, the thief disrupted the electricity and the machine became less productive from that time on. There is a prime suspect who has an alibi between midnight and 3AM (he was seen at a bar), so the police have a special interest in determining if the break-in occurred before midnight or after 3, since the suspect has no alibi for that time interval. We have data on Mus-Brush production every two minutes from 8PM until 5AM. Our job is to decide whether there is enough evidence to argue that the break-in occurred between 12 and 3, thus getting the suspect off the hook.

Figure 6. Competing models task: Mus-Brush Company (Rubin et al., 2006)

In contrast to these first two styles of tasks, Zieffler et al. (2007) used a population of quantitative data as a basis for the null model (see Figure 7) and asked students to make decisions about whether a certain educational outcome observed in a sample of data (change in mean test score) was due to chance or not. This was part of a multipart task that asked students to first imagine different possible samples for the population and sketch a hypothetical graph of the distribution of these samples.

Shown below is a graph of scores from many sections of students who have taken this course and exam. For this population, the average score is 74.



A random sample of 50 students in the class this year, given the exact same exam, had a mean exam score for of 78.

1. Do you think that the teacher can say that this year's students did better on average than what would be expected? Explain.
2. Do you think this higher sample average score could just be due to chance?

Figure 7. Competing models task modified from Zieffler et al. (2007)

3.4. ANALYSIS OF THE THREE TYPES OF TASKS

Table 1 illustrates how each of the three types of tasks shown incorporates the essential components of IIR that have been described in the research literature (see Section 2.3).

These tasks can be used in an interview, or on a written assessment to capture students' informal inferential reasoning, or can be embedded in classroom activities designed to promote the development of IIR. Each task may be used to reveal the extent to which students have integrated their prior knowledge about foundational concepts. They challenge students to make judgments and predictions about a population without the use of formal statistical methodology. Lastly, by having students explain their reasoning, usually more than once during the task, they elicit the articulation of students' arguments and justifications for their predictions and judgments.

Table 1. Specification of how each type of task incorporates the three components of IIR

TYPE OF TASK	IIR COMPONENT Make judgments or predictions	 Use or integrate prior knowledge	 Articulate evidence-based arguments
Estimate and draw a population graph	Predict characteristics of a population (shape, center, spread) that are represented in a student-constructed graph	Bring in intuitive or previously learned knowledge and language to predict the characteristics of a population (e.g., idea of shape, words like skewed)	Requires an explanation of how the characteristics of the population graph were chosen
Compare two samples of data	Judge whether there is a difference between two populations; based on similarities or differences in samples of data.	Bring in intuitive or previously learned knowledge and language to compare two samples of data (e.g., between- and within-groups variation)	Requires an explanation of why students determined whether or not there is a difference in the two populations
Judge between two competing models	Judge whether sample data provide more support for one model than another	Bring in intuitive or previously learned knowledge and language to judge between two competing models (e.g., sampling variability, chance variation)	Requires an explanation of why students chose one model over the competing model

These tasks (or parallel versions of the tasks) could be given to students at multiple times throughout a course or unit of instruction to examine how students' reasoning develops. This would allow instructors to examine how students use their informal knowledge and informal reasoning to draw conclusions and make inferences as they experience instruction related to informal or formal methods of statistical inference. This assessment could be done formatively which would allow for more opportunity for feedback to students, which in turn, would create more learning and research opportunities.

4. SUMMARY AND IMPLICATIONS FOR RESEARCH AND TEACHING

In this paper, we set out to answer two main questions regarding informal inferential reasoning. First, what are the components of a framework needed to support research on informal inferential reasoning? Drawing on the research literature, we proposed a working definition of IIR that comprises three components: (1) making judgments, claims, or predictions about populations based on samples, but not using formal statistical procedures and methods (e.g., p-value, *t* tests); (2) drawing on, utilizing, and integrating prior knowledge to the extent that this knowledge is available; and (3) articulating evidence-based arguments for the judgments, claims, and predictions about populations based on samples.

The second aim of this paper was to identify the types of tasks suggested by this framework for the study of informal inferential reasoning and its development? We have proposed two approaches that might be taken by researchers in studying IIR. We have also suggested the types of tasks that might be helpful in these types of studies, as well as provided concrete examples of such tasks.

The IIR framework (see Section 2.3) and suggested tasks (see Section 3) can be useful in further research studies in various ways. First, by referring to a common working definition in future studies, researchers may better be able to build on and connect their work to previous and subsequent studies. Second, using common tasks allows for comparisons across different instructional settings and groups of students. Third, these tasks have value in both teaching and research settings as they may be used in class activities to promote student reasoning and to challenge students to explain and articulate their understanding and rationale for making inferences, which could be studied to see how students' reasoning changes during the activity.

Researchers could also draw on the proposed IIR framework to help analyze students' responses to tasks designed to elicit IIR. Drawing on the first component of the framework, a researcher might ask whether the student made reasonable inferences about one or more populations based on one or more samples. For example, for the task in Figure 1 (predicting characteristics of a population from a sample), a reasonable inference might be that the center (mean) of the population is near the value that appears to be at the center of the sample, and the variability in the population is likely to be greater than the variation displayed in the sample.

Drawing on the second component of the framework, a researcher might ask how the student used and integrated informal knowledge (e.g., everyday knowledge of the problem context, prior knowledge about statistical concepts, real world knowledge and experience, and statistical language) in making inferences. Another question might be whether the use of problem context has impeded, or over-ridden, the use of data in making inferences. For example, using the same task, an integrated response might incorporate ideas of random sample as being representative of the population; ideas of distribution (e.g., shape, center, variation); and real world knowledge of the problem-context (e.g., test scores, college students' study habits). Another question of interest might be how heavily the student depends on prior knowledge of statistics (previously learned concepts) and how much the student depends on his or her knowledge of the world (or experience), a balance that may change over the course of instruction.

Drawing on the third component of the framework, a researcher might ask how the student has used evidence to support his/her arguments in making inferences, and also, how well the evidence used supported the inferences made. For example, using the same task, the response should include data-based explanation for why the student chose a particular population distribution (e.g., the population will likely have a mean near 71 because the sample had a mean of 71.3 and this sample was drawn randomly from the population so it should be representative).

Although the IIR framework proposed may be useful in studying the development of IIR (e.g., during an activity, or over a unit of instruction, an entire course, or even a curriculum) there is not yet a developmental model of how IIR develops from the earliest and most informal stage to transitioning to formal statistical reasoning. A variety of theories of developmental growth exist that could be used to underpin a "developmental" framework for IIR. One such cognitive model of learning is the Structure of Observed Learning Outcome (SOLO) based developmental framework (see Pegg, 2003). Based on the research presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), Reading (2007) suggested that a two-cycle SOLO-based framework be considered for describing the cognitive growth students experience when reasoning about statistical inference. The first cycle would involve reasoning about underlying concepts, including naïve inference that is not chance related. The second cycle would involve using these underlying concepts in a more "formal" way that incorporates reasoning about chance events.

The authors note that a working definition for IIR is a definition in progress. They hope that over time others will contribute to refining and updating this definition as more information is gained about the nature and development of students' informal inferential reasoning. There is a need for more research to explore the role of foundational concepts, data sets and problem contexts, and technology tools in helping students to reason informally, and then formally, about statistical inference. There are many unanswered questions about the best sequence of ideas and activities and the role of these in making the transition from informal to formal methods of statistical inference.

ACKNOWLEDGEMENTS

We appreciate the support of a grant from the National Science Foundation (NSF CCLI – 0535912).

REFERENCES

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal*, *3*(2), 64-83.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(2)_Bakker.pdf]

Bakker, A., Derry, J., & Konold, C. (2006). Technology to support diagrammatic reasoning about center and variation. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D4_BAKK.pdf]

Bakker, A. & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D1_BENZ.pdf]

Bransford J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Carver, R. (2006, August). *Ambiguity intolerance: An impediment to inferential reasoning?* Paper presented at the Joint Statistics Meetings, Seattle, WA.

Cobb, P., & McClain, K. (2004). Principles of Instructional Design for Supporting the Development of Students' Statistical Reasoning. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 375-396). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Cobb, P., McClain, K., & Gravemeijer, K. P. E (2003). Learning about statistical covariation. *Cognition and Instruction, 21*, 1-78.

Cobb, P., Yackel, E., & Wood, T. (1992). A constructivist alternative to the representational view of mind in mathematics education. *Journal of Research in Mathematics Education, 23*(1), 23-33.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997-1003.

Confrey, J. (1990). A review of the research on students' conceptions in mathematics, science, and programming. *Review of Research in Education, 16*, 3-56.

delMas, R., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education, 7*(3). [Online: http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm]

Driver, R., Newton, P., & Osborne, J. (2000), Establishing the Norms of Scientific Argumentation in Classrooms, *Science Education, 84*, 287-312.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove (UK): Lawrence Erlbaum Associates Ltd.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology, 5,* 75-98.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching*. New York: Springer.

Garfield, J., delMas, R., & Chance, B. (2007). Using students' informal notions of variability to develop an understanding of formal measures of variability. In M. Lovett and P. Shah (Eds.), *Thinking with Data (Proceedings of the 33rd Carnegie Symposium on Cognition)* (pp. 117-147). New York: Erlbaum.

Gravemeijer, K., & Doorman, K. (1999). Context problems in realistic mathematics education: A calculus course example. *Educational Studies in Mathematics, 39*(1/3), 111-129.

Jones, G. A., Langrall, C. E., Mooney, E. S., & Thornton, C. A. (2005). Models of development in statistical reasoning. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 97-117). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kahneman, D., Slovic, S., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Klahr, D., Fay, A., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology, 25,* 111-146.

Kuhn, D. (1991). *The skills of argument*. Cambridge, UK: Cambridge University Press.

Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development, 60*(4, Serial No. 245).

Lipson, K. (2003). The role of the sampling distribution in understanding statistical inference. *Mathematics Education Research Journal, 15*(3), 270-287.

Mack, N. K. (1995). Confounding whole-number and fraction concepts when building on informal knowledge. *Journal of Research in Mathematics Education, 26*(5), 422-441.

Means, M. L., & Voss, J. F. (1996). Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and Instruction, 14*(2), 139-178.

Miller-Jones, D. (1991). Informal reasoning in inner-city children. In J. F. Voss, D. N. Perkins, and J. Segal (Eds.), *Informal reasoning and education* (pp. 107-130). Hillsdale, NJ: Lawrence Erlbaum Associates.

Moore, D.S. (1990). Uncertainty. In L.A. Steen (Ed.), *On the shoulders of giants* (pp. 95-173). Washington, DC: National Academy Press.

Moore, D.S. (2004). *The basic practice of statistics* (3rd ed.). New York: W. H. Freeman.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Nickerson, R., Perkins, D. N., & Smith, E. (1985). *The teaching of thinking*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Papert, S., & Harel, I. (Eds.) (1991). *Constructionism.* Norwood, NJ: Ablex Publishing.

Pegg, J. (2003). Assessment in mathematics: A developmental approach. In J. Royer (Ed.), *Mathematical Cognition* (pp. 227-259). Greenwich, CT: Information Age Publishing.

Perkins, D. N. (1985a). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology, 77*(5), 562-571.

Perkins, D. N. (1985b). Reasoning as imagination. *Interchange, 16*(1), 14-26.

Perkins, D. N. (1989). Reasoning as it is and could be. In D. Topping, D. Crowell, & V. Kobayashi (Eds.), *Thinking: The third international conference* (pp. 175-194). Hillsdale, NJ: Lawrence Erlbaum Associates.

Perkins, D. N., Bushey, B., & Farady, M. (1986). *Learning to reason* (Final report for grant no. NIE-G-83_0028). Cambridge, MA: Harvard Graduate School of Education.

Perkins, D. N., Farady, M., & Bushey, B. (1991). Everyday reasoning and the roots of intelligence. In J. F. Voss, D. N. Perkins, and J. Segal (Eds.), *Informal reasoning and education* (pp. 83-105). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pfannkuch, M. (2005). Probability and statistical inference: How can teachers enable learners to make the connection? In G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 267-294). New York: Springer.

Pfannkuch, M. (2006). Informal inferential reasoning. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/6A2_PFAN.pdf]

Quilici, J. L., & Mayer, R. E. (2002). Teaching students to recognize structural similarities between statistics word problems. *Applied Cognitive Psychology, 16,* 325-342.

Reading, C. (2007, August). *Cognitive development of reasoning about inference*. Discussant reaction presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), University of Warwick, UK.

Rossman, A. (2007, August). *A statistician's view on the concept of inferential reasoning.* Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), University of Warwick, UK.

Rubin, A., Hammerman, J., & Konold, C. (2006). Exploring informal inference with interactive visualization software. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D3_RUBI.pdf]

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching, 41*(5), 513-536.

Sadler, T. D., & Zeidler, D. L. (2004). The significance of content knowledge for informal reasoning regarding socioscientific issues: Applying genetics knowledge to genetic engineering issues. *Science Education, 89,* 71-93.

Saldanha, L. A., & Thompson, P. W. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*(3), 257-270.

Saldanha, L. A., & Thompson, P. W. (2006). Investigating statistical unusualness in the context of resampling: Students exploring connections between sampling distributions

and statistical inference. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/6A3_SALD.pdf]

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49,* 31-57.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology, 32,* 102-119.

Schauble, L., & Glaser, R. (1990). Scientific thinking in children and adults. In D. Kuhn (Ed.), *Developmental perspectives on teaching and learning thinking skills* (pp. 9-27). Basel, Switzerland: Karger.

Schoenfeld, A. H. (1982). Measures of problem-solving performance and of problem-solving instruction. *Journal for Research in Mathematics Education, 13*(1), 31-49.

Schoenfeld, A. H. (1991). On mathematics as sense-making: An informal attack on the unfortunate divorce of formal and informal mathematics. In J. F. Voss, D. N. Perkins, and J. Segal (Eds.), *Informal reasoning and education* (pp. vii-xvii). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 484-494.

Schwarz, D. L., Sears, D., & Chang, J. (2007). Reconsidering prior knowledge. In M. Lovett and P. Shah (Eds.), *Proceedings of the 33rd Carnegie Symposium on Cognition: Thinking with Data*. Mahweh, NJ: Erlbaum.

Smith, J. P., diSessa, A. A., & Roshelle, J. (1993/1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2), 115-163.

Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*, 98-113.

Stohl, H. & Tarr, J. E. (2002). Developing notions of inference using probability simulation tools. *Journal of Mathematical Behavior, 21*, 319-337.

Tarr, J. E., Stohl Lee, H., & Rider, R. (2006). When data and chance collide: Drawing inferences from simulation data. In G. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth NCTM yearbook* (pp. 139-150). Reston, VA: National Council of Teachers of Mathematics.

Thompson, P., Saldanha, L., & Liu, Y. (2004, April). *Why statistical inference is hard to understand*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego.

Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

van Eemeren, F. H., Grootendorst, R., Henkemans, F. S., Blair, J. A., Johnson, R. H., Krabbe, E. C. W., et al. (1996). *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Mahweh, NJ: Erlbaum.

Voss, J. F., Blais, J., Means, M. L., Greene, T. R., & Ahwesh, E. (1986). Informal reasoning and subject matter knowledge in the solving of economics problems by naïve and novice individuals. *Cognition and Instruction, 3*(3), 269-302.

Voss, J. F., Perkins, D. N., & Segal, J. W. (1991). Preface. In J. F. Voss, D. N. Perkins, and J. Segal (Eds.), *Informal reasoning and education* (pp. vii-xvii). Hillsdale, NJ: Lawrence Erlbaum Associates.

Walton, D. N. (1989). *Informal logic: A handbook for critical argumentation*. Cambridge, UK: Cambridge University Press.

Watson, J. M. (2004). Developing reasoning about samples. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 277–294). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Watson, J. M., & Moritz, J. B. (1999). The beginnings of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*, 145-168.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-265.

Zieffler, A., delMas, R., Garfield, J., & Gould, R. (2007, August). *Studying the development of college students' reasoning about statistical inference.* Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), University of Warwick, UK.

ANDREW ZIEFFLER
Educational Psychology
Room 250 EdSciB
4101
56 E River Road
Minneapolis, MN 55455

EXPLORING BEGINNING INFERENCE WITH NOVICE GRADE 7 STUDENTS

JANE M. WATSON
University of Tasmania
Jane.Watson@utas.edu.au

ABSTRACT

This study documented efforts to facilitate ideas of beginning inference in novice grade 7 students. A design experiment allowed modified teaching opportunities in light of observation of components of a framework adapted from that developed by Pfannkuch for teaching informal inference with box plots. Box plots were replaced by hat plots, a feature available with the software TinkerPlots^{TM}. Data in TinkerPlots files were analyzed on four occasions and observed responses to tasks were categorized using a hierarchical model. The observed outcomes provided evidence of change in students' appreciation of beginning inference over the four sessions. Suggestions for change are made for the use of the framework in association with the intervention and the software to enhance understanding of beginning inference.

Keywords: *Statistics education research; Hat plots; Informal inference; Middle school students; TinkerPlots*

1. LITERATURE REVIEW

The word "inference" can convey many shades of meaning, depending on the context in which it is used and the adjectives that may be placed in front of it. A dictionary such as *Chambers* (Kirkpatrick, 1983) suggests an inference is "that which is inferred or deduced: the act of drawing a conclusion from premises: consequence: conclusion" (p. 644). For a statistician it is the premises that are important. For David Moore (1991) statistical inference is the "formal methods for drawing conclusions from data taking into account the effects of randomization and other chance variation" (p. 330). The parts of Moore's definition that preclude it from most of the school curriculum are the "formal" methods, the "randomization," and the "chance" variation to the extent that it relies on formal probability. Within the statistics education community, the phrase "formal inference" is usually used synonymously with "statistical inference" in Moore's sense. This usage leads to attempts to define "informal inference." Certainly at the middle school level replacing "formal methods" with "informal methods" is appropriate as the mathematics required for formal methods is not available to students. The question of what constitute appropriate informal methods is then a matter for debate among statistics educators. Although randomization may be introduced to middle school students through chance devices such as coins and dice, the link to the need for random methods generally may be difficult for students to grasp and even more difficult to implement during data collection activities. Chance variation is an idea that is accessible in an intuitive fashion but nearly impossible to connect to numerical values that would express degrees of uncertainty.

Statistics Education Research Journal, 7(2), 59-82, http://www.stat.auckland.ac.nz/serj

This raises the question, *What aspects of the process of drawing conclusions from data can be kept and what left out, and still use the phrase "informal inference"?* Rubin, Hammerman, and Konold (2006) give their summary of important ingredients of informal inference as including properties of aggregates related to centers and variation, sample size, controlling for bias through sampling method, and tendency (acknowledging uncertainty). Watson and Moritz (1999) use the phrase "beginning inference" to describe the comparison of two finite data sets because the activity is a precursor to the use of *t* tests by statisticians. The context is similar but the premises differ. Although the techniques available fall under the heading of exploratory data analysis, the requirement to make a decision about the difference between the two groups takes the activity into the realm of beginning or informal inference. The admission of uncertainty about the decision reached, although based on intuition and maybe graphs rather than numbers, mirrors the process involved with formal inference.

The National Council of Teachers of Mathematics' (NCTM) *Principles and Standards for School Mathematics* (2000) use the words "develop and evaluate inferences and predictions" in the Data Analysis and Probability Standard throughout the document for all ages, with increasingly sophisticated and rigorous criteria used for the techniques that "are based on data" (pp. 48-51). Hence, the use of the word "inference" on its own does not necessarily imply "formal inference." Although they do not use the phrase, the descriptions of the NCTM would imply agreement with the definition of "informal inferential reasoning" by Ben-Zvi, Gil, and Apel (2007):

> Informal Inferential Reasoning (IIR) refers to the cognitive activities involved in informally drawing conclusions or making predictions about "some wider universe" from patterns, representations, statistical measures and statistical models of random samples, while attending to the strength and limitations of the sampling and the drawn inferences. (p. 2)

There are many components in this definition and it is questionable whether a unit on beginning inference in most middle school programs would be able to cover them with a high level of sophistication.

The context of the study reported here included the provision and introduction of the software TinkerPlots^{TM} (Konold & Miller, 2005). This allowed for the consideration of the use of the software in relation to the goals of informal inference. TinkerPlots was designed using a bottom-up constructivist approach for the development of students' data handling and analysis skills (Konold, 2007). Students build representations using tools and the drag-and-drop facility, rather than choosing a "type" of plot that is ready-made. One particular tool of relevance to this research is the hat plot. It is a simpler version of the box plot. In its basic default form the hat consists of a crown that covers the middle 50% of the data with brims on either side covering the 25% of data at each extreme. A hat plot for a small data set is shown in Figure 1. The median is not represented in the hat plot although it can be marked on the horizontal axis where the data are stacked. A hat plot initially appears in conjunction with the data set it represents, providing visual reinforcement.

The similarity of a hat plot to a box plot suggests that the research of Pfannkuch (2005, 2006a, 2006b) on informal inference with box plots in the final years of school may be instructive for analyzing the data in this study. Pfannkuch devised a model for the development of inferential reasoning based on box plots. The model evolved from Pfannkuch's intensive planning with a classroom teacher before teaching informal inference based on box plot representations, observations in the classroom during teaching, analysis of the data, and various discussions with statisticians and statistics educators. The eight elements of reasoning in the model are (i) *hypothesis generation*,

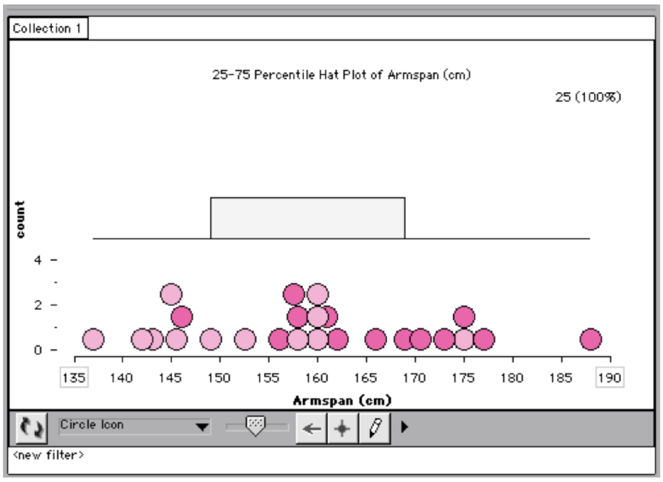


Figure 1. Hat Plot for a data set

(ii) *summary* of the plot, (iii) considering *shift* of box plots, (iv) comparing the *signal* in the central 50% of data, (v) considering *spread*, (vi) discussing aspects of *sampling*, (vii) considering *explanatory* aspects of the *context*, and (viii) considering *individual cases* (Pfannkuch, 2006b). Two other moderating elements were also described relating to weighing up *evidence* for evaluation and to using concept *referents* that lay behind the box plots, which were the objects of the lessons. As these were considered by Pfannkuch to be included in the main eight elements (2006b, p. 33), they are not included in this study. Evaluation and adaptation of the model was based on observations of teachers and their students in classrooms, assessment tasks, and associated student outcomes. One of the difficulties observed in the classroom was the abstract nature of the box plot, which was presented without the display of data values, putting extra pressure on students to recall the proportional reasoning associated with interpreting the plot. Pfannkuch's conclusions recommended that the data and box plot should appear together.

In relation to the use of box plots, Bakker's (2004) research on using computer tools with middle school students led to the questioning of the introduction of box plots to middle school students (Bakker, Biehler, & Konold, 2005). The features of box plots that cause difficulties for students include (i) the lack of ability to examine individual values, (ii) the different form of presentation from other graphical representations, (iii) the non-intuitive nature of the median, and (iv) the visual presentation of density rather than frequency. Based on the five-number summary (minimum, maximum, median, and first and third quartiles) the box plot is an intermediate summary of data between a graph and a single summary statistic such as the arithmetic mean. As such it may hide some interesting features of a data set. Bakker et al. recommended that the introduction of box plots be delayed until later in the middle school years and that more time be spent relating box plots to the data sets they represent. These recommendations support the possibility that hat plots as available in TinkerPlots may provide a viable alternative to box plots for middle school students. Other recent research using TinkerPlots with both students (Friel, O'Connor, & Mamer, 2006) and teachers (Rubin & Hammerman, 2006) illustrates aspects of Pfannkuch's model adapted for use with the different visual presentations provided by TinkerPlots. Friel et al. note the use of dividers, box plots, and reference lines to assist students in looking at data sets as aggregates in much the same way that hat plots are used in this study. Rubin and Hammerman use bins and dividers to consider slices of data sets for similar purposes. The significant aspect of visualization in both studies, enforcing attention to distribution, highlights the four elements of Pfannkuch's model related to Spread, Shift, Signal, and Summary. Using a framework for assessing enabling aspects of software applied to statistical environments, Fitzallen (2007) found that TinkerPlots (i) was accessible and easy to use, (ii) assisted recall of knowledge and allowed data to be represented in multiple forms, (iii) facilitated transfer between mathematical expression and natural language, (iv) provided extended memory when organizing or reorganising data, (v) provided multiple entry points for abstraction of concepts, and (vi) provided visual representations for both interpretation and expression (p. 24).

In relation to the model for observing the development of inferential reasoning with box plots by Pfannkuch (2006b), Konold et al. (2002) reported on observations of student descriptions of stacked dot plots that focused on "modal clumps." These clumps of data values occupied central locations in data sets and usually divided the distribution of values into three parts. The percentage of values in each part depended on the overall shape of the distribution. It appeared natural for the students to "see" the distributions in three parts: low, middle, and high. The modal clumps were often described in conjunction with a generalized average that was wider than a specific value. These observations also link to the display presented by a hat plot, with the natural splitting of the data set into three parts. Whether these possibilities for clustering assist in making decisions about differences in two data sets has received little research attention. Cobb (1999), however, observed students in a classroom study begin to consider the "hills" apparent in two data sets and how these might determine a difference in two groups. Watson and Moritz (1999) also found students using shape to distinguish between two classes' test results where the classes were of a different size.

2. CONTEXT AND RESEARCH QUESTIONS

The context within which the opportunity to conduct this study arose was a larger research project providing professional learning in mathematics for rural middle school teachers in an Australian state. The professional learning was aimed at providing teachers with the chance to plan and develop skills to assist their students with the mathematical foundations both for quantitative literacy as a life skill and for moving on to higher mathematics courses. The schools in the project were considered culturally isolated from the capital city and their students were performing on average below the national benchmarks for literacy and numeracy. TinkerPlots software was provided as part of the project to all schools and hence some professional learning sessions were based around it. Schools or teachers could elect to be involved with the researchers in case studies as part of the larger project. This provided the setting for addressing the following research questions:

- What are the observed learning outcomes for a class of grade 7 students, in relation to beginning inference in a learning environment supported by a learning and assessment framework, professional learning for the teacher, and a software package for data handling?
- How do the learning outcomes suggest changes to the framework, the implementation, or the interaction with software?

3. METHODOLOGY

The implementation of the study fits into a design experiment model of educational research as outlined for example by Cobb, Confrey, diSessa, Lehrer, and Schauble (2003) and The Design-Based Research Collective (2003). Five criteria derived from these authors are satisfied by the study. First, the design intertwined the starting framework (or theory) for development of informal inference based on box plots with the classroom intervention employing TinkerPlots software that provided the hat plot as an interpretation tool. Second, the design was interactive in that four data collection sessions took place with adaptation to plans and analysis between sessions through ongoing discussion among the researcher, teacher, and a teacher-researcher (T-R). Third, the study was highly interventionist in the provision of a T-R to mould activity when the teacher appeared insecure and in the expectation of student participation in an environment that was unfamiliar to them (TinkerPlots). In terms of initial expectations there were both successes and failures that required adjustment. Fourth, a variety of data sources were connected to document the design and intervention, including student outcomes, with descriptive accounts and structural analysis of students' TinkerPlots output. Fifth, the modification to the teaching and classroom environment was related to the theoretical framework as a way to document student outcomes and make recommendations for future intervention.

To expand on these criteria in order to present the results in context, the following sub-sections outline the background in terms of the participants, including the preliminary professional learning of the classroom teacher, the basic context for each of the four teaching sessions, the nature of student output and its analysis in terms of the adapted Pfannkuch framework, and the other sources that contributed to the modifications made to theory and practice.

3.1. PARTICIPANTS' BACKGROUND

The study was based on a class of approximately 15 grade 7 students (aged 12-13 years) in a rural school (grades K to 10) in Australia. Not all of the students were present at all sessions. This was a convenience sample dictated by a larger research project. The class teacher, Jenny (pseudonym), was in charge of all sessions except the last. A research associate who was a trained teacher (the T-R), assisted the author in preliminary sessions with teachers and was present at all student classes, taking detailed notes, assisting with use of TinkerPlots, and occasionally teaching.

Before the four sessions where data were collected, Jenny attended a professional learning day with the author and T-R where TinkerPlots was introduced as a tool to assist in drawing informal inferences from data sets. The aims included

(i) learning to enter data sets into TinkerPlots,

(ii) using a student-created data set and TinkerPlots to discuss differences in data over time and how to determine if a "real" change has occurred,

(iii) comparing data sets to decide if they are different,

(iv) considering the issue of sample size when making a decision, and

(v) becoming aware of the populations that samples represent.

Jenny supported other teachers experiencing similar activities and planned data collection and teaching sessions for her own students. Her aim was to address the five points.

3.2. LESSON CONTEXTS

Session 1 Prior to the session, pulse rates were collected from the middle-school students (grades 5 to 8) after a "Jump Rope for Heart" activity. Jenny's grade 7 students had explored TinkerPlots on their own earlier and the initial two-hour session with the heart-rate data was designed by Jenny and observed by the T-R. Jenny planned for the students to enter their class data in the TinkerPlots data cards and then create plots to describe the variation in the observed pulse rates and what might be considered typical of the group. After a discussion of their observations and suggested reasons for differences (proposed inferences), the students were to measure their resting heart rates, record these for the class, and enter the data onto their existing data cards. The objective then was to create another plot in TinkerPlots and compare it with the first in order to speculate about differences in the two sets of data. The final aim was for the students to write a report in a text box in TinkerPlots about the inferences/conclusions drawn.

Between Sessions 1 and 2, the author, Jenny and the T-R discussed more specifically the aims of informal inference in relation to the elements of the adapted Pfannkuch model. Although the initial professional learning for Jenny and for the other teachers where Jenny assisted had covered this material generally (but not the eight specific points) it was felt necessary to be more explicit in discussing them. These are summarized in Table 1 as a Beginning Inference Framework. It was intended that Jenny be aware of them but she was not asked to "teach" them to her class explicitly as a list. She agreed that this was her understanding.

Table 1. Eight elements of a Beginning Inference Framework (adapted from Pfannkuch, 2006b)

Element	Description
Hypothesis Generation	Reasons about trends (e.g., differences)
Summary	Summarizes the data using the graphs and averages produced in TinkerPlots
Shift	Compares one hat plot with the other/s referring to change (shift)
Signal	Refers to (and compares) information from the middle 50% of the data
Spread	Refers to (and compares) spread/densities locally and globally
Sampling	Considers sampling issues such as size and relation to population
Explanatory/ Context	Shows understanding of context, whether findings make sense, and alternative explanations
Individual Case	Considers possible outliers and other interesting individual cases

Session 2 Some of the students assisted fellow middle-school students with a similar activity the day after Session 1, reinforcing their software skills. Jenny then devoted another day with her class to reviewing the heart-rate task and gaining an appreciation for the value of the TinkerPlots hat plot. This was to be a very structured session because of the students' lack of background with hat plots and percentages generally. The aims reported by Jenny were to link these two ideas, as well as to focus on the ranges, means, medians and scales (horizontal axis) of the two heart-rate graphs. Again the plan was for students to record their observations about differences in the two sets of data using the new tools and reasons for these differences (informal inference).

Session 3 To reinforce the skills and understanding achieved, several weeks later Jenny collected data on arm-span length from all of the 58 middle-school students. This was organized by Jenny outside of the class time and she prepared a color-coded handout for students to use when entering their data. The plan for this session was jointly devised by Jenny and the T-R. The plan was to create a "class dictionary" of terms on the whiteboard to reinforce the previous sessions. The class would then decide on variables in the data set, enter the data into TinkerPlots, produce plots to explore the data, and make conjectures about any differences they observed. Again students were to write a final report in a text box in TinkerPlots, explaining their evidence to support their conclusions (informal inferences) related to the arm-span data set.

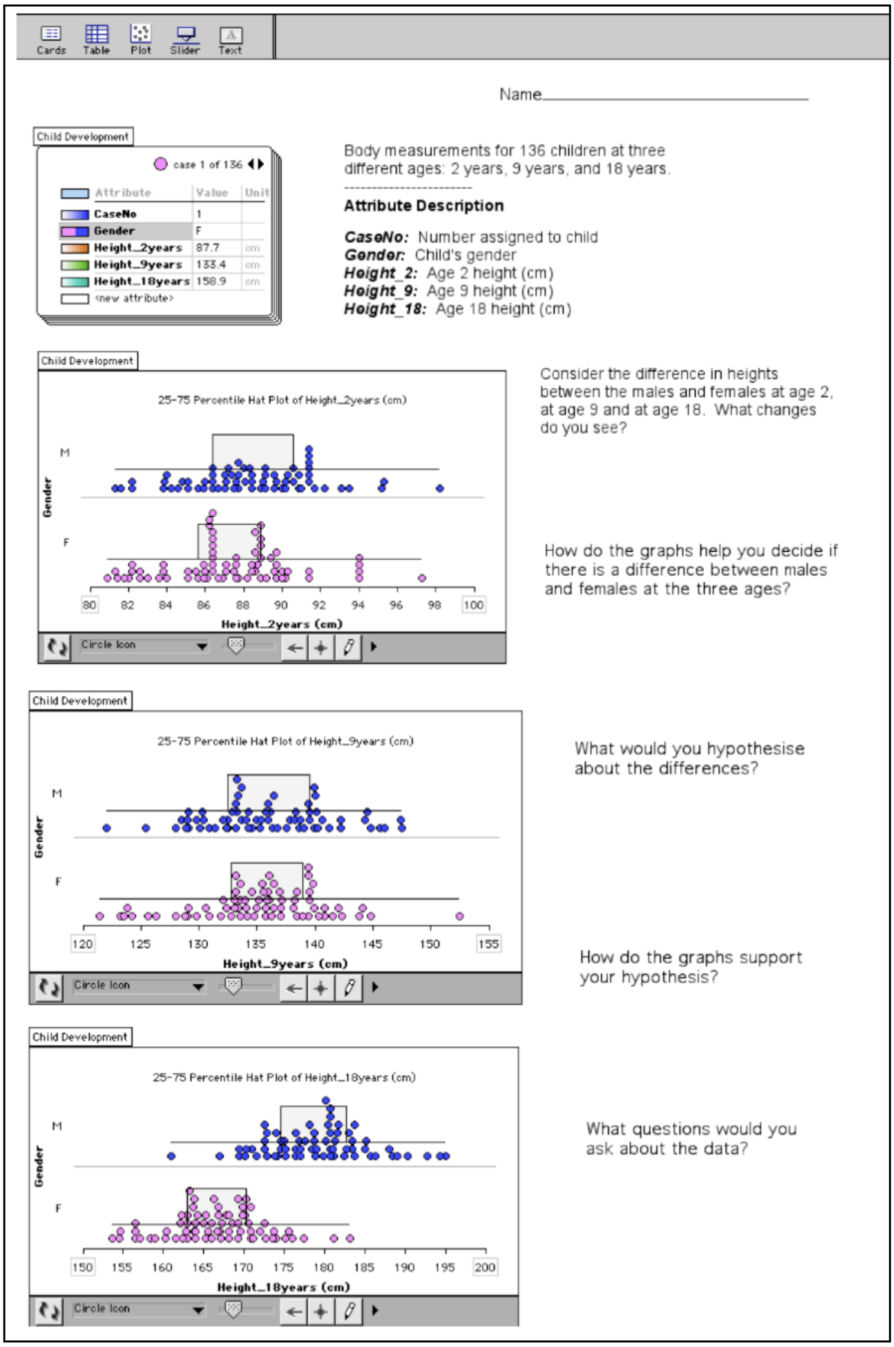
Session 4 At the beginning of the following school year, approximately 3½ months after Session 3, the T-R returned to the school for a final session with the students, now in grade 8 with a different teacher. The T-R taught the entire session with another teacher-researcher observing and the classroom teacher assisting a student with special needs. Because of the time gap, the session was planned to start with a review of the arm-span measurement activity, including a review of the terms used such as *x*-axis, *y*-axis, bins, stacking, mean, median, hat plot, brim and crown, reference lines, dividers, percentages, range, spread, and clustering. The T-R created several TinkerPlots graphs from the data using a data show with students grouped around a table. Two new terms were to be introduced for specific discussion and definition: hypothesis and evidence.

Students were then each to be provided with the TinkerPlots file as shown in Figure 2 and asked to complete the questions presented in the text boxes. The data were from a data set provided with the TinkerPlots software (Child Development.tp) and the data set was reduced to case number, gender, height at age 2 years, height at age 9 years, and height at age 18 years. Three graphs were produced with hat plots and the questions were designed to encourage students to make and support informal inferences. After answering the questions presented in Figure 2, students were to be given the opportunity to explore the data set further by changing the graphs in any way they liked, especially if it would tell the story better.

3.3. OUTPUT AND ANALYSIS

The model suggested by Pfannkuch (2006b) to describe the aspects of informal inference associated with the introduction of box plots was adapted for both the teaching and the assessment of student work completed using TinkerPlots. The eight elements of the Beginning Inference Framework outlined in Table 1 cover well what would be expected in the context of a TinkerPlots exploration for middle-school students. Analysis was based on TinkerPlots files created by students during each session and saved at the ends of the sessions. The first step in the analysis was to determine which of the eight aspects of the Beginning Inference Framework were displayed in each response for each student. Grids were created with 8 cells for each of the 18 students who participated in at least one of the four sessions. A “tick” in a cell indicated that a graph was produced that appeared to address an aspect of Pfannkuch’s model but that no text was entered into a text box. Text in a cell provided a summary of text box comments made by the student. A clustering procedure (Miles & Huberman, 1994, p. 248) was used to group together comments related to the elements of informal inference displayed by each student at the end of each session.

At this point the clusters of elements of informal inference were assessed using the Structure of Observed Learning Outcomes (SOLO) model devised and employed by psychologists Biggs and Collis (1982) and mathematics educator Pegg (2002). In their developmental and assessment framework, it is the manner in which the elements are combined that shows increased sophistication and complexity in relation to achieving a desired outcome. In this study, the desired outcome was a description of an informal inference taking into account the components of the Beginning Inference Framework. As a simplified example, when first presented with a data set, a student may only be interested in observing her own place in the data set, or the largest and smallest values in the set. This would be considered a *unistructural*, single-element response within the



Name______________________________

Attribute	Value	Unit
CaseNo	1	
Gender	F	
Height_2years	87.7	cm
Height_9years	133.4	cm
Height_18years	158.9	cm
<new attribute>		

Body measurements for 136 children at three different ages: 2 years, 9 years, and 18 years.

-------------------------

Attribute Description

CaseNo: Number assigned to child
Gender: Child's gender
Height_2: Age 2 height (cm)
Height_9: Age 9 height (cm)
Height_18: Age 18 height (cm)

Consider the difference in heights between the males and females at age 2, at age 9 and at age 18. What changes do you see?

How do the graphs help you decide if there is a difference between males and females at the three ages?

What would you hypothesise about the differences?

How do the graphs support your hypothesis?

What questions would you ask about the data?

Figure 2. Tasks for Session 4

context of the overall expectation of a beginning informal inference task. Another student might look at a graph, plot the mean and report the average value for the data set. Again the response would be considered unistructural. A student who put these two elements together, or perhaps added another comment on spread, would likely be considered to produce a *multistructural* response, especially if the comments were made in a sequential fashion without being specifically related to each other. Taking a number of the elements and combining them to make a related coherent argument in terms of informal inference would be considered a *relational* response. For example, a relational response might observe the shape of a data set, with clustering in one part, and spread in another, discuss an individual outlier, and compare these features with those of another data set hypothesizing a difference or sameness between them. No response or an idiosyncratic response is called *prestructural*.

3.4. OTHER SOURCES

The classroom interactions are described based upon the written notes of the T-R, long discussions of the researcher and T-R, and short meetings or telephone discussions with the teacher. One longer meeting of the teacher, T-R, and researcher was held after the first session. These sources are intertwined with evaluations of happenings in relation to the positive or negative impacts on the classroom atmosphere and student observed outcomes.

4. RESULTS

4.1. SESSION 1 – HEART-RATE DATA

Because students had explored TinkerPlots in a previous lesson, the expectation was that all students would open data cards to enter the data. After this there was, however, some confusion on the relative importance of name and heart rate and where to position them on the plots. Because of the way the data were entered on the TinkerPlots data cards, the names were paramount. Students fixed on these because they knew themselves and their classmates, and because “name” was the initial attribute on the data card. Almost all students created either a value plot or a plot with the same information on heart rates represented as circle icons. Some ordered these from lowest to highest heart rate and many noted the person with the highest rate and the person or persons with the lowest.

At this point the T-R intervened and asked all students to open a text box and write what their graphs told them about the heart rates. No one had yet stacked the dots on a dot plot. When asked about the “average” heart rate, most students looked at the data table and selected the most common value, 120 beats per minute. Some looked at highest and lowest values and commented on fitness. The resting heart rates were then measured. The T-R questioned the students about where the new data would go, what they were as attributes, and why the same data cards were used to record resting values. Then students had to make decisions about how to display the new data. When required, the T-R encouraged the students to create two graphs, one for each of “active” and “resting” rates. They did this in various fashions, often with value bars or bins, or pie charts within bins.

Of the 13 students present at the first session, all except two considered Individual Case aspects of the data related to the people in the class and largest or smallest measurements; for example, “Resting… 1. When resting Peter has a high heart rate. 2. When resting nick [s.,] callum and bianca have a low heart rate.” Because of the enormous interest in the activity and differences in individual values for the class members, Jenny did not discourage the discussion of these points. Four of the students followed up on this by including Explanatory/Context comments in their text boxes in TinkerPlots, for example, “running 1. nick is the most unfit. 2. krystal is the fittest. But krystal may not have been running as much as nick so her heart may not have been working as hard as nicks.” Context was explicitly mentioned by three students in relation to the fitness of the class. These comments could be considered as embryonic hypotheses but only for the context of the data collected for this class, not for any wider population.

As a Summary statement, average of some sort was mentioned by five students with two noting the most common value, while others used the mean as recorded on their graphs by the software: “running … the average heart rate is around 140 because the computer has added the heart rates up and divided that by the amo[u]nt of people in the class.” Spread was discussed by four students, one providing the following comments: “1. There is a huge difference between highest and lowest on both of the graphs. 2. It makes it different depending on the range of the scale.” The difference in scale for the two graphs (60-80 vs 100-176) was the focus of concern in class discussion, with students again saying it was not easy to make comparisons. Near the end of the lesson students were taught how to make the scales of two graphs the same but few had time to do so.

Four students produced graphs but no accompanying comments in text boxes. Two students used bins to organize the data but did not explain the representation. Those who created scatterplots of “active” and “rest” rates either made no comment or said the graph was not easy to interpret. Only two students presented hat plots for the two data sets with the same scale. These students had no other graphs in their files and made no comments in text boxes. They may have deleted earlier graphical representations and the absence of text indicates the two students may have followed the procedures with little understanding. In Session 1 no students entered comments in text boxes related to Shift, Signal, or Sampling. All students created at least two plots in TinkerPlots and eight were considered to produce Unistructural responses in their text boxes as shown in the quotes in this section. One student might be considered to be providing a Multistructural response in considering three aspects of the data. Four students who entered no comments in text boxes were considered prestructural in their observed learning output. The only move beyond the data was seen in attempts to explain the data within the classroom context. In considering the 104 cells in the analysis matrix for the 13 students and 8 elements of the Beginning Inference Framework, 31% were annotated indicating students had produced graphs or text.

The output shown in Figure 3 is typical of the student interest in the heart-rate values for the class. The plots illustrate the students’ use of bins, scale, and value plots.

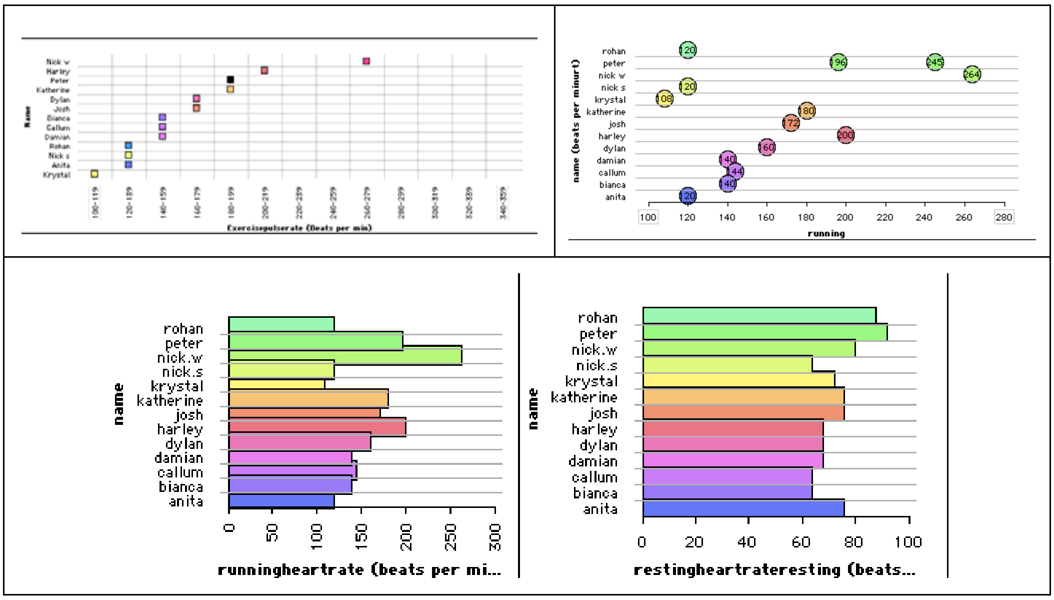
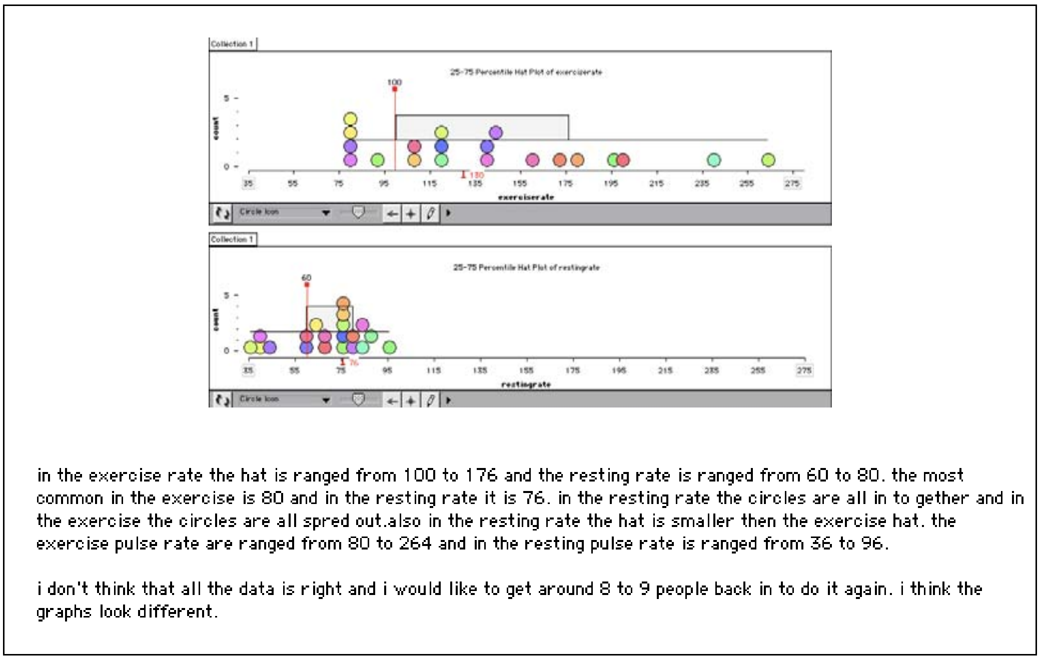


Figure 3. *Representations of individual heart rates for each of the students in the class created in Session 1*

4.2. SESSION 2 – HEART-RATE DATA

Much of the time in Session 2 was spent by Jenny explaining the concepts related to hat plots and percentage and hence there was little opportunity to experiment with different representations of the heart-rate data sets and compare them using hat plots. Besides hat plots, discussion focused on the ranges of the two data sets, their means and medians, and the necessity to have the same scale for each graph. All students, except one, created hat plots for the two data sets. Student output in TinkerPlots files was collected at this point for analysis. All students created two similarly scaled graphs and included at least one hat plot. Three students included no text comments. The other eight students made comments about the Spread of the data, some of which were related to the central 50% of data covered by the crown of the hat. No one commented on Individual Case values; this may have been because they had studied the data earlier or because of the relatively structured nature of the lesson. Only one student commented on the Explanatory/Context in terms of exercise and two noted Sampling issues, related to the validity of the data recorded. The T-R reported that in overheard conversations a few students were unwilling to suggest a "difference" in heart rates at the two times because there was some overlap in the data. The graphs and text from one student are shown in Figure 4. The student's response was considered to employ the elements related to Shift in commenting on the different ranges, to Signal in noting and comparing the range of the circles "together" and "all spread out," and to aspects of Sampling in questioning the validity of some values and wanting some to repeat the trial.



in the exercise rate the hat is ranged from 100 to 176 and the resting rate is ranged from 60 to 80. the most common in the exercise is 80 and in the resting rate it is 76. in the resting rate the circles are all in to gether and in the exercise the circles are all spred out.also in the resting rate the hat is smaller then the exercise hat. the exercise pulse rate are ranged from 80 to 264 and in the resting pulse rate is ranged from 36 to 96.

i don't think that all the data is right and i would like to get around 8 to 9 people back in to do it again. i think the graphs look different.

Figure 4. Student extract from Session 2

The differences in the elements of the Beginning Inference Framework addressed in Session 2 compared to Session 1 are quite striking, reflecting the teacher input of the session. Of the observed student outcomes in the TinkerPlot files, Summary of the graph is the main overlapping feature, whereas there was a strong shift from considering individual cases to considering spread and the middle 50% of the data in relation to the hat plot in Session 2. Only one student commented specifically on an average value (Summary), whereas five noted Shifts in the data sets, three commented specifically on the centre 50% (Signal), and eight discussed Spread. Only one commented on Individual cases, and two made comments on the Sample (as in Figure 4).

In Session 2 three students made no comment in the text boxes, two made single Unistructural responses in relation to spread, and six made two or three comments of a Multistructural nature. Of the 88 cells in the analysis matrix for the 11 students and 8 elements of the Beginning Inference Framework, 43% were annotated indicating either a graph or text had been produced. The purpose of the session on developing the appreciation for hat plots had focused student attention on ways of distinguishing two data sets, a precursor to larger aspects of informal inference.

At the end of Session 2, the T-R led a discussion about the validity of the data collected. Although they had not talked about it before, students could offer suggestions about why the data could be in error. These included miscounting or not concentrating. To remedy the situation students suggested having someone else take the heart rate, checking the method used, or having the same person take all of the measurements.

4.3. SESSION 3 – ARM-SPAN DATA

The T-R returned and helped Jenny teach a session where the class analyzed the data collected on the arm spans of the middle-school students (see Section 3.2). At the beginning of the class, the T-R assisted with recording the class list of terms. It was decided to enter four attributes on the cards: name, grade, arm span, and sex. Students were then asked to think about different ways of exploring the data, with questions like “What does it tell you?”, “Is it meaningful?” They were encouraged to use the terms on the whiteboard as starting points for their investigations. They were asked to keep all graphs and write a text box about what each one told them. The students were actively involved and discussed what they were doing with each other. Most students had created at least three graphs by recess. At recess, students were provided with morning tea and sat around a large table, discussing different ways of looking at data. They summarized aspects of the data, for example the differences between some grades (e.g., 5 and 6, or 6 and 7) but not others (e.g., 7 and 8). This introduced the topic of growth spurts, which led to a discussion of differences between boys and girls. Although the means and medians for boys and girls were nearly the same, the spread for boys was much greater. The tallest and shortest grade 7 boys talked about their advantages and disadvantages. Talk about spread led to making the point that one needs to know more than the mean or median of a data set. Then the issue of sample size was raised because there were different numbers in each grade and of each gender. The most difficult question discussed was related to representativeness, for example when asked, “Which ‘group’ was the data we have collected true for?” and in particular how would they tell if the data were “true” for students in grade 5 to 8 in Tasmania, someone suggested measuring the arm span of every grade 5 to 8 student in Tasmania. The concept of sampling a few schools appeared to be difficult for all of the students.

After recess students went back to the computer lab and were asked to write a Final Report: to choose a comparison of interest, and using the class dictionary on the whiteboard as a guide to write their report. Different tools were to be used, including hat plots, dividers, and percentages, and students were asked to recommend the ones they thought were most useful in making the comparison. The saved files from Session 3 were collected for analysis.

The spread of responses across the Pfannkuch elements was much greater than in the first two sessions. Three of the students created separate files for the Final Reports and most appreciated the task of summarizing their work and providing some sort of evaluation of their use of TinkerPlots. The only aspect not covered was that of Sampling, a topic that the T-R reported discussing at recess and one that she felt the students found difficult to comprehend. Due to the lack of an initial discussion of a wider population and hypothesis, this is not surprising. Three students considered the context of student growth with grade, which may have been a result of the discussion at recess. Four students appeared to address Hypothesis Generation for differences across grades but these were locally-based observations. It was mainly the use of language that separated these from the Explanatory/Context comments. Except for one student who did not write in text boxes, all students discussed Summaries of the data and their plots, including averages and observed differences. The word "shift" was not used with the box plots but change was addressed by six students. Spread, with nine specific comments mainly about range, and the Signal in the central 50% of the data, with six specific comments, were considered at about the same level as for Session 2 but as not all students used hat plots in Session 3, the references were generally encouraging. Irrelevant information not related to arm span but to the number of boys or girls in groups was provided by six students, two of whom had not been present during Sessions 1 and 2. All students made some comment on Individual Cases, presumably related to their encounter with a new data set where they knew the participants.

Although some students struggled with their informal inferences, they appeared to be appreciating the process involved. The following two comments by one student are associated with the graphs displayed in Figure 5. The response was judged to be a Relational response in the context of this study.

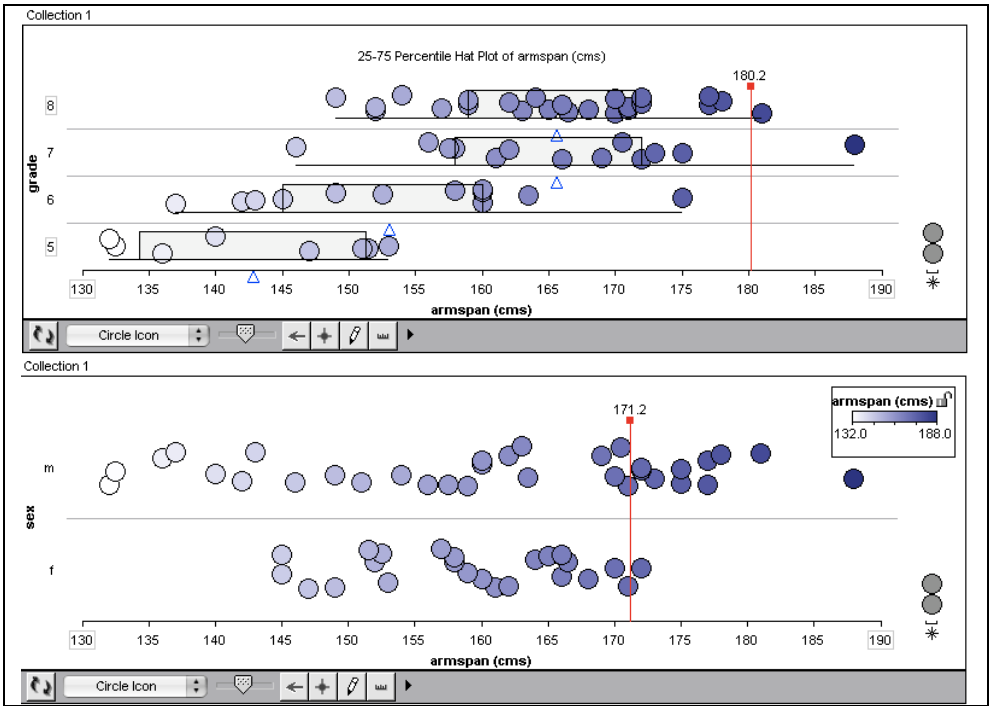


Figure 5. Figures related to student comments for arm span data

> [I] com[pared] the arm span and grades. there was an 11 cent[i]meter difference between the mean of grade 5 and 6, and a 12cm difference between the mean of grades 6 and 7, but between grades 7 and 8 there was almost no difference at all in the mean. This tells me that the students do most of their growing in grades 6 and 7. [Figure 5, top]
>
> I compared the armspan with boys and girls. The range of the boys results was much m[o]re spread out[.] They went from 132cms to 188.1. The girls went from 145cms to 172cms. I think this is because girls are usually smaller than boys. [Figure 5, bottom]

Another student used hat plots for the comparison of the grade 5 and grade 8 students. The descriptive account of what the student found out about the data is quite extensive (the student uses "hat" for the crown of the hat). The text box entry relates to Figure 6:

> i find the difference between the age and the size of the arm span interesting. i used the hat because it shows me that the hat is covering 50% of the dots. in the grade 8 the hat is around 159.0 to 171.0 and in the grade 5 it[']s around 136.1 and 151.6. i think the hat is easy to use because it shows me where the most armspans are ranged from. i used the mean and the median because it shows me that the average in the mean in the grade 8 is 165.6 and in the grade 5 it is 145.1 but the median in the grade 8 is 164.9 and in the grade 5 its 147.0. in the range of the grade 5 the smallest is conner with 132 and the highest is bryce with 163.5. but in the grade 8 the lowest is kasey with 149 and the highest is tim with 181.

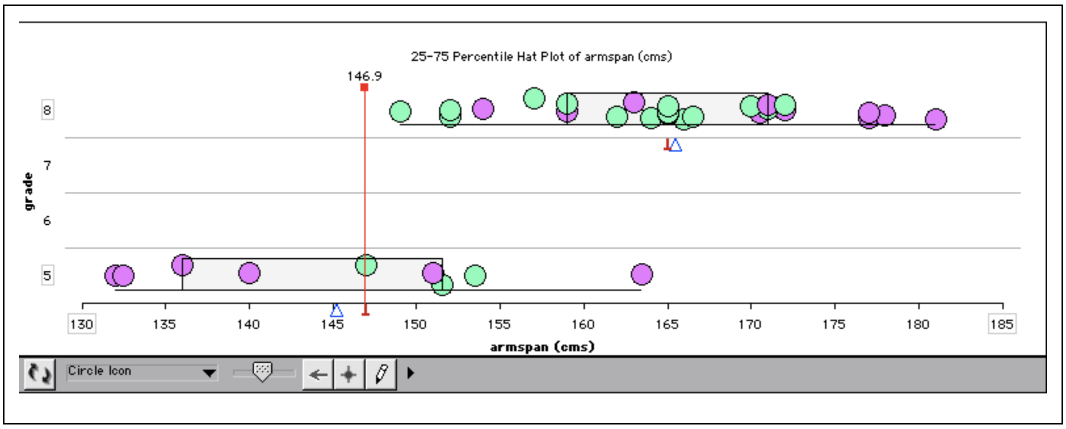


Figure 6. Student plot from Session 3 using hat plots

The student then produced the graph in Figure 7 and explained the difference for grades 5 and 8 in terms of percentages related to a central "average." This is an interesting comparison but the criterion for determining the central average is not explained: "with the div[i]der and the % i can see that 60% of the grade 8s are the average and 40% is higher th[a]n the average. in the grade 5 44% is the average and 56% is lower th[a]n the average." Overall this student's response was judged Relational in the context of this study.

Acknowledging that students during Session 3 were asked to keep all of their output and to prepare a Final Report, it is not surprising that a wider coverage of Beginning Inference Framework elements occurred. All students volunteered that they liked using TinkerPlots generally but some still reported that they found hat plots difficult. It is likely that this difficulty is related to their previous mathematical background, especially their understanding of percentages.

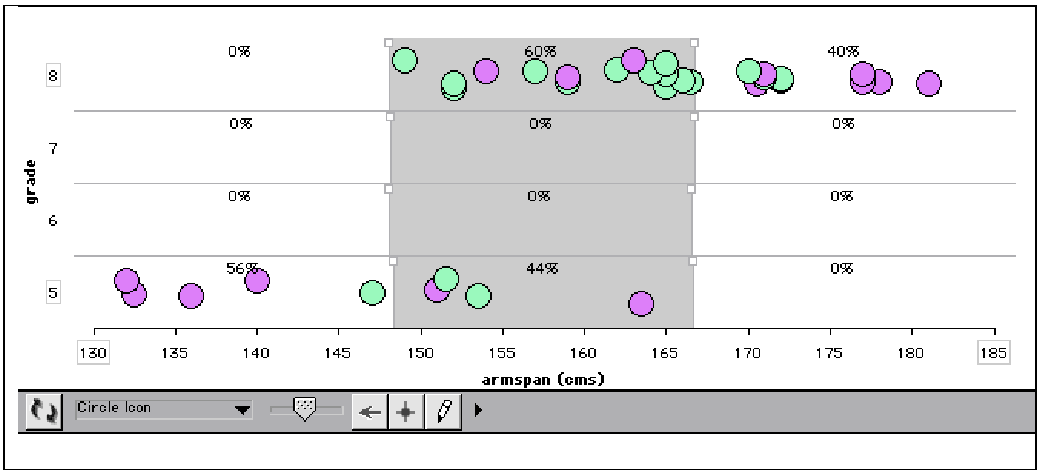


Figure 7. Student plot from Session 3 using dividers

In this session, where students experienced the most freedom in completing their analyses, only the aspect of Sampling did not receive any attention. With a new data set students were again interested in Individual Cases. The consideration of Explanatory/Context was often related to the informal Hypothesis Generation about the data set. The linking of consideration of context/hypotheses and the Summary of plots and Spread appeared to result in Relational responses by six students, whereas six others appeared as Multistructural in sequencing Summaries of the plots, Spread, and Individual Cases. Only one student did not fill in at least some text boxes. Of the 104 matrix cells in the Beginning Inference Framework for the 13 students and 8 elements, 57% indicated that students had produced a graph or written text.

4.4. SESSION 4 – GROWTH DATA

In the session 3½ months later at the start of the following school year, the T-R led a discussion of terminology, including the specific terms hypothesis and evidence. Students contributed to the discussion, which concluded that hypotheses were the questions they wanted to answer and evidence related to how they knew something, for example to support their hypotheses. Students were then given TinkerPlots files to work on.

The specific questions in the text boxes (see Figure 2) encouraged the students to consider the eight components of the Beginning Inference Framework. Two girls answered the questions on print-out rather than using text boxes. A few students, however, continued to struggle, leaving some text boxes unfilled. Whether this was due to lack of understanding, lack of literacy skills, or general unwillingness to complete the task is unknown.

The T-R reported that the conversation among the students was of a higher quality than the written responses, as students were reluctant to compose expressions in written words. The difficulty with literacy skills is apparent in some of the responses. The relationship between the questions about hypotheses and descriptions of the graphs was confusing for some students. As well, meeting a much larger data set for the first time made it difficult for some students to decide if a relatively small difference in appearance constituted a genuine “difference.”

As the graphs in Figure 2 were not altered by the students in this part of Session 4, student responses from the text boxes from three students for the five questions are presented in Table 2.

Table 2. Selected extracts from three students to the questions presented in Session 4

Consider the difference in heights between the males and females at age 2, at age 9 and at age 18. What changes do you see?	**How do the graphs help you decide if there is a difference between males and females at the three ages?**	**What would you hypothesize about the differences?**	**How do the graphs support your hypothesis?**	**What questions would you ask about the data?**
They are getting taller. At 18 years the boys are about 10 cms taller. [Student 1]	the hat plot shows the difference between boys and girls at the 3 ages.	up till about 11 years boys and girls are a similar height. when get to 18 years the boys get taller than the girls.	the graphs show that the hat plots separate when the boys and girls turn 18.	what is the arverage hight for both boys and girls at 18 years. why do girls reach their hights early before boys
the spread becomes different the older they get but the real growing happens in the 9 years between 9 and 18. [Student 2]	the hat plots are the main help they make it easier to compare the girls and the boys.	boys in general are higher, girls are smaller but when yopunger in children they appear the near same.	I actually got the hypothesis from the graphs so they support what I have been saying considerably.	maybe do a graph showing the boys/girls at 2 and a bar underneath at 18 so you can see how they have grown in this way you would need names near each dot.
when the females are younger there are more near the end of the line. when the get old the start to go near the start of the line. 2 years-the males are spred out more then the girlsin the hat plots and the hat plots over lap each other. 9 years- theres not much difference between the middle 50% and most of the hat plots over lap each other. 18 years-there a little difference between the middel 50%. the hat plots start to seperate from each other. [Student 3]	the hats help me see the middel 50%(crown) and the 1st and last 25%(brimms) between the males and the females. when there stacked it shows me show how many people are the height. it also shows the range og people in the hat.	that the males would be taller then the females when they were younger and older but in the middel age the females would be taller.	it show that the males are the tallest in the younger and older age and the middel age a female is taller then the males. at 2 years the males are taller then the females. at 9 years the males was the tallest but one of the females was taller then the males. at 18 years the males was the tallest of all.	is this infomation we have ture or is it guessed? have we got the real mesurments or is the mesurments rounded off? what are the range of males at the age of 9? what percentage of girls are under the height of 100 cm? what is the average hight for a 14 year old? what range can a 14 year old grow to?

The text in Table 2 has not been edited, reflecting exactly what was written in text boxes. Not all responses were as articulate as these. In a couple of cases other students produced similar responses to these indicating that discussion and sharing of ideas took place among class members. Student 1 was considered to have addressed four of the elements in the Beginning Inference Model: Hypothesis Generation (up until 11 about the same then boys getting taller), Summary (at 18, boys about 10 cm taller), Shift (hat plots separate at 18), and Explanatory/Context (why do girls reach heights before boys?). This was judged a Multistructural response as a sequence of comments in these four areas. Student 2 also addressed four elements in the model: Hypothesis Generation (boys in general higher, when young nearly the same), Summary (hat plots easier to compare boys and girls), Spread (different as they get older), Explanatory/Context (real growing from 9 to 18). These comments were considered to be more integrated and hence Relational in nature. Student 3 addressed all of the elements in the model: Hypothesis Generation (males taller when younger and older; females in the middle), Summary (range and heights [frequencies]), Shift (at 18, hats separate), Signal (at 9 years, middles the same), Spread (shows range, hats overlap), Sampling (is information true or guessed?), Explanatory/Context (what range at 14), and Individual Case (at 9, one female taller than all males). This response was judged to be Relational.

Of importance in the responses in Table 2 is the continued use of the article “the,” apparently indicating that the students were considering their hypotheses in relation to the data set presented rather than a larger population of children, which would be referred to without the article. Hence, it appears that although some steps had been taken to use the language of hypotheses across groups, it does not represent a genuine attempt at generalization to a much wider population.

When given the opportunity to change the plots in the file to tell the story of the data set better, 12 of the students made some changes to at least one of the graphs in Figure 2. Of these, five made no additional written comments, for example changing to square or “people” icons and un-stacking the data, sometimes recombining the males and females in the data set, or using a small number of bins. Others commented on the fusing of icons or using of dividers as helping them to see crucial aspects of the data for telling the story. Three boys scaled the three graphs from 80 to 200, one with un-stacked square icons and no hats. One of the other representations is shown in Figure 8. Two of the comments show that the task was appreciated in terms of helping to justify the responses provided earlier in the text boxes.

> Student 4: this shows them all at the 18 year old range, this is particu[lar]ly us[e]ful for showing how the people have grown over the time period[.] there proba[b]ly could have been a graph made in the 12 year old range to bridge that huge gap between 9 and 18.

> Student 5: the graphs that i have just made makes it [ha]rder to disting[ui]sh the h[e]ights but easier to compare them to their selves 16 [yrs] down the track. this ma[d]e it easier to compare [their] 2yr olds to the 18 yr olds and so forth. but doing this also made the hats harder to distinguish.

Two girls produced scattergraphs with height at age 18 on the horizontal axis and height at age 9 on the vertical axis. One is shown in Figure 9 and is color-keyed by gender. As the students had had virtually no experience with interpreting scattergraphs,

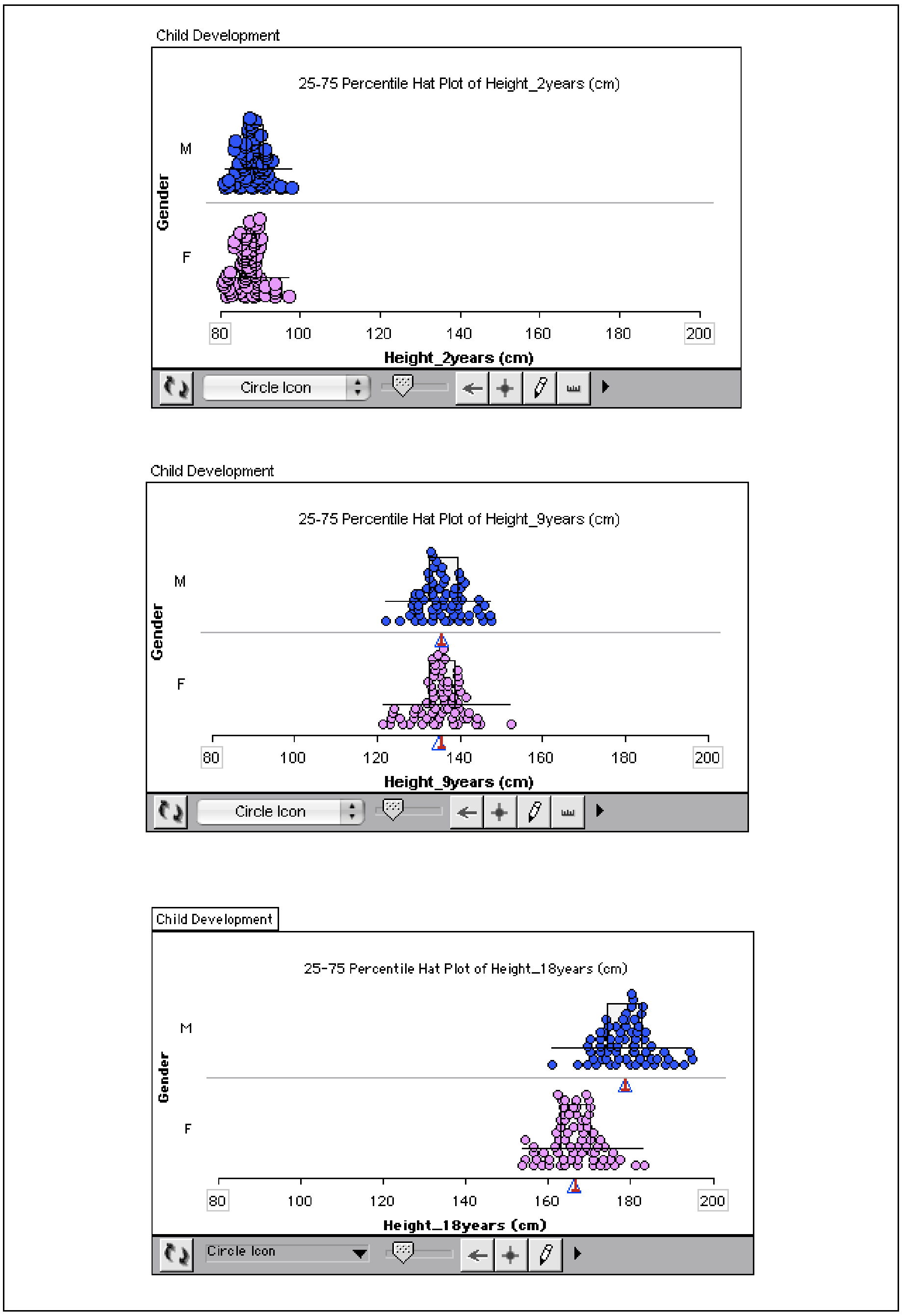


Figure 8. Scaling of height data by students in Session 4

the comments partly missed the point. “This graph shows that as you get older the more you grow. [A]nd that males are mainly taller then females but some are the same.” The attempt, however, demonstrates a growing appreciation of the task of generating hypotheses from the data.

In the final session, the structured questions scaffolded Relational responses in 11 out of 15 cases with the other 4 being Multistructural. The interest lay in how students expressed themselves. Overall these responses showed what the students had picked up from the previous sessions and their beliefs in the usefulness of dividers, reference lines,

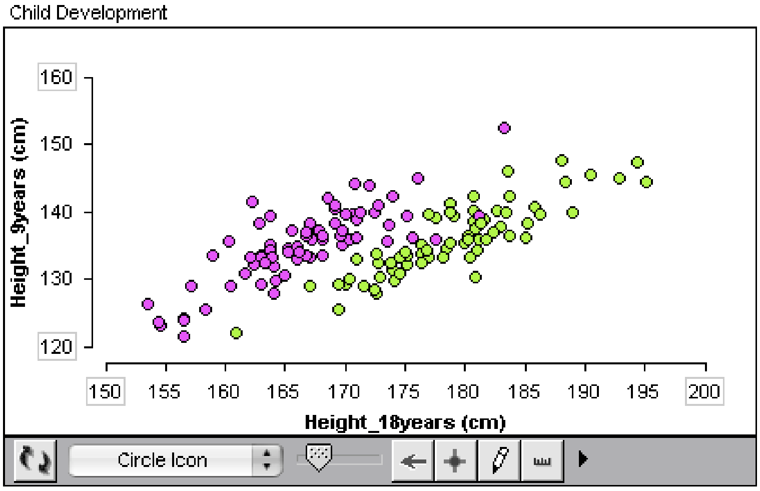


Figure 9. Scatterplot for height data

and scale in telling stories about data sets. Of the 120 matrix cells in the Beginning Inference Framework for the 15 students and 8 elements of the framework, 67% indicated text responses. The fact that text boxes were provided with questions may have supported this increased percentage.

4.5. SUMMARY

The clustering of students' observed outcomes within each session is summarized in the previous four sections. Because of the different increasing levels of facility with TinkerPlots and the different degrees of scaffolding across the teaching sessions, it is somewhat difficult to compare the structural complexity of students' responses written in text boxes over time. The levels reported in Table 3, however, show higher levels of observed responses across sessions. Jenny felt that in terms of student motivation and application to the tasks these observations were indicative of higher achievement than usual for her class. Further, a perusal of the comments for the three students reported in Table 2 shows that appropriate language was being taken up by these students.

Table 3. Summary of structural complexity of observed responses over four sessions

Session	1[a]	2[b]	3[c]	4[d]
No. of students in class	13	11	13	15
No text response (Prestructural)	4	3	1	0
Uni-structural	8	2	0	0
Multi-structural	1	6	6	4
Relational	0	0	6	11

[a]Heart-rate data entry. [b]Heart-rate hat plots. [c]Arm-span data entry and analysis. [d]Interpretation of height/age/gender data.

Eight of the students were present for all four sessions where data were collected. A summary of their levels of text box entries is presented in Table 4. The improved observed performance of these students reflected that of the overall group. That these students could take up many of the elements of the Beginning Inference Framework and use them with guidance represents a first step to informal inference. The comments of Students B and C for Session 4 are among those presented in Table 2.

Table 4. Levels of observed response for eight students who were present for all four teaching sessions

Student	Session 1[a]	Session 2[b]	Session 3[c]	Session 4[d]
Student A	U	U	M	M
Student B	U	M	R	R
Student C	U	P (no text)	R	R
Student D	U	M	R	M
Student E	U	U	M	R
Student F	M	M	M	R
Student G	P (no text)	M	M	R
Student H	U	P (no text)	M	R

Note. P = Prestructural, U = Uni-structural, M = Multi-structural, R = Relational
[a]Heart-rate data entry. [b]Heart-rate hat plots. [c]Arm-span data entry and analysis. [d]Interpretation of height/age/gender data.

5. DISCUSSION

5.1. STUDENT CHANGE IN RELATION TO BEGINNING INFERENCE

In relation to the research questions that guided this case study, the first concerns the observed learning outcomes for the students in relation to beginning inference. Using the Beginning Inference Framework adapted from Pfannkuch (2006b), the students used the facilities offered by the TinkerPlots software to address elements associated with Summary, Shift, Signal, Spread, and Individual Cases. At the grade 7 level it was more difficult to distinguish the elements of Hypothesis Generation and Explanatory/Context as context was essential to both elements and to the students' language of questioning and speculating. Sampling was the most difficult element for students to assimilate. It is likely that this difficulty arose from the order in which the lessons proceeded, the limited life experiences of the rural students, and the delayed emphasis on the rather abstract nature of the sample-population relationship. As noted, the T-R found from her interaction with the students that they were having difficulty with it.

In terms of the observed learning outcomes as assessed from the TinkerPlots files created by the students (with two exceptions being hand written for the final session), the SOLO levels of observed outcomes, as well as the percentage of elements addressed, increased with the sessions. The scaffolding of the classroom experiences as well as the nature of the questions in the final session undoubtedly contributed to this increase, but this was indeed the aim. The long-term nature of this understanding past Session 4 is impossible to determine from the study.

According to Jenny, the students had had no prior experience with graphing software and their graphing experience appeared restricted to bar charts, and in a couple of instances pie charts. Students, however, adapted easily to the drag-and-drop features of TinkerPlots and most, after the first session, were able to select appropriate variables to place on the axes. In terms of the goal of comparing two (or more) groups, the majority of students appreciated the value of hat plots. In the early stages they had some difficulty in judging whether "a difference" existed at all, even when there was very little overlap of two data sets, whereas later some students were willing to claim as genuine differences what appeared to be quite small differences. The use of the hats, both created and interpreted by students, proved helpful to some students in describing shifts in data sets, although the language employed was often quite colloquial (see Table 2). A few students at the end still preferred to put the data in bins but they could explain, for example to the T-R, what the hats represented. Students participated well in the class discussion but did not enjoy writing comments in the text boxes in TinkerPlots to record their reasoning. The written comments were not of the same caliber as the notes recorded by the T-R during the sessions. In two instances in the final session girls wrote their thoughts on a hard copy of the TinkerPlots output rather than using the text boxes. The hand-written text was much more extensive than anything presented in a text box in the earlier sessions by these two students. This and the observation of the literacy and typing levels of students lead to the suggestion that a second order of difficulty is introduced for many students when asked to express ideas not only in words (for example on paper) but also in typed text in a text box. This leads to the conclusion that the levels of response observed are likely to be minimum levels for these students.

The observed interaction of the learning framework, the teacher's preparation, and the software is important in evaluating the students' observed outcomes and suggesting changes. The teacher and the T-R were introduced formally to the Beginning Inference Framework after Session 1, when it was decided by Jenny to continue working with her students. The T-R implemented the aspects in her planning and at various points she raised them with the students. At no point, however, were students provided with notes or handouts suggesting they consider each component of the model. Due to the limited number of teaching sessions, not a great degree of reinforcement was possible. The T-R expressed some surprise, however, at what the students remembered over the summer period (3½ months) between Sessions 3 and 4. It is clear from the analysis that in a structured setting the students could apply many of the components of the model and some could relate them together in meaningful ways. That the students had used the TinkerPlots representations to assist in making their comments is evident from the explicit phrases in the text boxes. Providing a simpler version of a box plot, accessible to these students, appeared to be of assistance to those who made Relational comments. As noted throughout the study, however, some students employed dividers, reference points, means, and medians to supplement or replace hat plots in reaching decisions about the data sets.

That the point reached by the students in this case study along an imaginary "informal inference" continuum was not very far past "beginning informal inference" is clear. They had, however, begun to address some of the criteria set by Rubin et al. (2006). The progress observed by the T-R, herself an experienced teacher, however, was considered extraordinary given her observation of students' starting points and the general background of schools in the larger project. The students certainly made progress and all except one claimed in a feedback form that they would use TinkerPlots again.

5.2. IMPLICATIONS FOR FUTURE RESEARCH AND IMPLEMENTATION

The second research question follows from the first in the light of the experiences during the four classroom sessions. The limitations of this case study are those of most educational research in less-than-ideal settings. The students were from a rural setting where formal education is not always valued; as is seen in some of the student extracts, literacy levels were low for grade 7. Eighteen different students were present in Jenny's class for at least one of the four sessions, whereas only eight were present for all four. Three areas for changes in future research and implementation are considered.

The use of the Beginning Inference Framework adapted from Pfannkuch (2006b) was very useful for most of the eight elements. The interaction of the elements Hypothesis Generation and Explanatory/Context, however, caused some difficulty in allocating students' observed responses to one category or the other. In explaining the context as they saw it, students sometimes used it as a foundation to suggest a hypothesis. In the future, increased emphasis on sampling and the relationship of samples to a population may assist students in being more explicit in generating hypotheses more generally than for the specific context, for example, of their class. The issues associated with using the framework for planning teaching and for assessment lead back to the general question of how many and which of the elements can or should be introduced over time during the middle years. For teachers who themselves have little experience with inference, there appear to be difficulties in absorbing the content to the extent of appreciating what their students experience as learners and hence in developing pedagogies that are appropriate. A large-scale study considering alternative orders of introduction of the elements might offer guidance in this area.

Although Jenny had experienced professional development before the teaching sessions, had assisted other teachers, had talked to the researcher and T-R, and was enthusiastic about introducing the software to her class, at times she lacked confidence in relation to the goals of an investigation in terms of beginning inference and turned to the T-R for guidance. In these situations the T-R was fairly directive in intervening with the class. Hence, it is likely that a more intensive initial introduction to the Beginning Inference Framework needs to precede a teacher's planning and implementation of similar activities. Teachers' enthusiasm for the software must be complemented continually with the goals of the middle school statistics curriculum. Whether a framework as complex as the one used in this study should be explicitly introduced to grade 7 students at the start of a unit of work is debatable. Further research would be needed to decide and the author has concern that it would be overwhelming to most grade 7 students, certainly those similar to the ones in this study.

The issue of literacy levels and some students' reluctance to complete text boxes in TinkerPlots suggests that future research could record student conversations while involved in the data analysis or that after the final session individual interviews could take place with the students involved. Although it may be assumed that in this computer age students are able to readily type into text boxes, this was not the case for most of this class. Future intervention with grade 7 students might allow a choice of how the conclusions of investigations are recorded. It is possible that students could print out their graphs and handwrite their reports and inferences on them. The requirement to assess observed learning outcomes can be met in several ways that may allow students the best chance of displaying their learning.

5.3. CONCLUSION

Using both TinkerPlots and the adapted Pfannkuch model (see Table 1) appears an appropriate and valuable resource for introducing beginning inference to middle school students. The software and model reinforce each other, explicitly in terms of the elements of Hypothesis Generation, Summary, Shift, Signal, Spread, and Individual Case. The other two elements of the model, Sampling and Explanatory/Context, although perhaps not directly linked to the software, provide cues to considering the other elements in relation to plots, for example to question sample size or make sense of alternative explanations. In terms of assessment, combining the eight elements of the Beginning Inference Framework with a developmental model from cognitive psychology appears to allow the documentation of students' observed outcomes. As such, it allows the combination of two important criteria: statistical appropriateness and structural complexity (Watson, 2006). Given the observation of students' initial difficulties in judging whether differences in two data sets were small or large, it would appear that the ease of creating representations and using TinkerPlots may contribute to developing intuitions about what might be considered "significant" or "meaningful" differences.

This study presents the hat plot as a viable alternative to the box plot, hopefully satisfying Bakker et al. (2005) in their request to postpone the latter's introduction. In the meantime, the introduction of hat plots appears to encourage the description of the shape of data as noted by Konold et al. (2002) and the comparison of multiple data sets without the necessity of means (e.g., Watson & Moritz, 1999). The progression toward beginning inference, although slow, appears well documented by the Beginning Inference Framework.

ACKNOWLEDGEMENTS

This study was funded by an Australian Research Council grant, No. LP0560543. The T-R for the study was Karen Wilson. Thanks go to Karen for vetting all descriptions in this report and providing a validity check on them.

REFERENCES

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools*. Utrecht: CD-ß Press, Center for Science and Mathematics Education.

Bakker, A., Biehler, R., & Konold, C. (2005). Should young students learn about box plots? In G. Burrill & M. Camden (Eds.), *Curricular Development in Statistics Education: International Association for Statistical Education 2004 Roundtable* (pp. 163-173). Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/rt04/4.2_Bakker_etal.pdf]

Ben-Zvi, D., Gil, E., & Apel, N. (2007, August). What is hidden beyond the data? Helping young students to reason and argue about some wider universe. Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5), University of Warwick, UK.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning*, *1*, 5-43.

Cobb, P., Confrey, J., deSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13.

Fitzallen, N. (2007). Evaluating data analysis software: The case of TinkerPlots. *Australian Primary Mathematics Classroom*, *12*(1), 23-28.

Friel, S. N., O'Connor, W., & Mamer, J. D. (2006). More than "Meanmedianmode" and a bar graph: What's needed to have a statistical conversation? In G. F. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth NCTM yearbook* (pp. 117-137). Reston, VA: National Council of Teachers of Mathematics.

Kirkpatrick, E. M. (1983). *Chambers 20th Century Dictionary* (New Ed.). Edinburgh: W & R Chambers Ltd.

Konold, C. (2007). Designing a Data Analysis Tool for Learners. In M. Lovett & P. Shah (Eds.), *Thinking with data: The 33rd Annual Carnegie Symposium on Cognition* (pp. 267-291). Hillside, NJ: Lawrence Erlbaum Associates.

Konold, C., & Miller, C. D. (2005). *TinkerPlots: Dynamic data exploration*. [Computer software] Emeryville, CA: Key Curriculum Press.

Konold, C., Robinson, A., Khalil, K., Pollatsek, A., Well, A., Wing, R., & Mayr, S. (2002). Students' use of modal clumps to summarize data. In B. Phillips (Ed.), *Developing a statistically literate society: Proceedings of the Sixth International Conference on Teaching Statistics*, Cape Town, South Africa. [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute. [Online: http://www.stat.auckland.ac.nz/~iase/publications/1/8b2_kono.pdf]

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Moore, D. S. (1991). *Statistics: Concepts and controversies* (3rd ed.). New York: W. H. Freeman.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

Pegg, J. E. (2002). Assessment in mathematics: A developmental approach. In J. M. Royer (Ed.), *Mathematical cognition* (pp. 227-259). Greenwich, CT: Information Age Publishing.

Pfannkuch, M. (2005). Probability and statistical inference: How can teachers enable learners to make the connection? In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 267-294). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pfannkuch, M. (2006a). Informal inferential reasoning. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute. [Online: http://www.stat.auckland.ac.nz/~iase/publications/17/6A2_PFAN.pdf]

Pfannkuch, M. (2006b). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal*, *5*(2), 27-45. [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ5(2)_Pfannkuch.pdf]

Rubin, A., & Hammerman, J. K. (2006). Understanding data through new software representations. In G. F. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth NCTM yearbook* (pp. 241-256). Reston, VA: National Council of Teachers of Mathematics.

Rubin, A., Hammerman, J. K. L., & Konold, C. (2006). Exploring informal inference with interactive visualization software. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute. [Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D3_RUBI.pdf]

The Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5-8.

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals*. Mahwah, NJ: Lawrence Erlbaum.

Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, *37*, 145-168.

JANE M. WATSON
Faculty of Education, University of Tasmania, Private Bag 66
Hobart, Tasmania 7001
Australia

DEVELOPING YOUNG STUDENTS' INFORMAL INFERENCE SKILLS IN DATA ANALYSIS

EFI PAPARISTODEMOU
European University Cyprus
e.paparistodemou@euc.ac.cy

MARIA MELETIOU-MAVROTHERIS
European University Cyprus
m.mavrotheris@euc.ac.cy

ABSTRACT

This paper focuses on developing students' informal inference skills, reporting on how a group of third grade students formulated and evaluated data-based inferences using the dynamic statistics data-visualization environment TinkerPlots™ (Konold & Miller, 2005), software specifically designed to meet the learning needs of students in the early grades. Children analyzed collected data using TinkerPlots as an investigation tool, and made a presentation of their findings to the whole school. Findings from the study support the view that statistics instruction can promote the development of learners' inferential reasoning at an early age, through an informal, data-based approach. They also suggest that the use of dynamic statistics software has the potential to enhance statistics instruction by making inferential reasoning accessible to young learners.

Keywords: *Statistics education research, Elementary education, TinkerPlots, Informal statistical inference*

1. OVERVIEW OF PROBLEM

Statistics, the science of learning from data, is divided into two main areas: descriptive statistics and inferential statistics. Descriptive statistics is the branch of statistics devoted to the organization, summarization, and presentation of data. It involves using tabular, graphical, and numerical techniques to analyse and describe a dataset. Inferential statistics, on the other hand, is intended to reach conclusions that extend beyond the immediate data, to deduce that observed patterns in the data at hand are also present in some broader context. It improves decision-making in a variety of real-world situations by providing tools that enable the drawing of causal inferences, or inferences to populations using sample-based evidence. Although statistical inference is the cornerstone of modern statistical concepts and methods, grasping the key ideas related to inferential statistics is a known area of difficulty for students (Green, 1982; Rubin, Bruce, & Tenney, 1990; Garfield & Ahlgren, 1998; Gordon & Gordon, 1992; Rubin, Hammerman, & Konold, 2006).

Traditionally, statistical inference is presented in the statistics classroom as a set of formal tests and procedures, through which information contained in sample data is used either to estimate the values of the respective population parameters (i.e., construct confidence intervals), or to check claims made regarding the values of population

Statistics Education Research Journal, 7(2), 83-106, http://www.stat.auckland.ac.nz/serj

parameters (i.e., perform hypothesis testing). Given the conceptual difficulties involved in understanding formal statistical inferential methods, introduction to statistical inference has traditionally been reserved for the high school or college level. At the lower levels of schooling, students' exposure to statistical concepts has been restricted to basic descriptive statistics. In recent years, however, leaders in mathematics education have advocated a much wider and deeper role for statistics in school mathematics (Shaughnessy, Ciancetta, Best, & Canada, 2004). It is now widely recognized that the foundations for statistical reasoning, including fundamental ideas of inferential statistics, should be laid in the earliest years of schooling rather than being reserved for high school or university studies (National Council of Teachers of Mathematics, 2000).

Advances of technology provide new tools and opportunities for the development of early statistical reasoning. Having such a set of tools widely available to young learners has the potential to give children access to advanced statistical topics including inferential statistics and the broader process of statistical investigation (Makar & Rubin, 2007), by removing computational barriers to inquiry. This leads to a shift in the focus of statistics instruction at the school level from learning statistical tools and procedures (e.g., graphical representations, numerical measures) towards more holistic, process-oriented approaches that go beyond data analysis techniques (Makar & Rubin). Statistics can be presented as an investigative process that involves formulating questions, collecting data, analyzing data, and drawing data-based conclusions and inferences (Guidelines for Assessment and Instruction in Statistical Education (GAISE) Report, 2005).

This paper reports on how a group of third-grade (8-year-old) students formulated and evaluated data-based inferences using the dynamic statistics data-visualization software TinkerPlots^{TM} (Konold & Miller, 2005), a statistical package specifically designed to meet the learning needs of students in the elementary and middle grades. We examine the role of the dynamic statistics tool in scaffolding and extending these young students' informal ideas of inference (Ben-Zvi, 2006).

2. LITERATURE REVIEW

Since formal statistical inference ideas and techniques are beyond the reach of young learners, an informal approach to statistical inference is necessary in the early years of schooling (Ben-Zvi, 2006). Developing students' informal ideas of inference is a topic of current interest to many statistics educators, who have acknowledged the fact that despite their difficulties with the formal methods of statistical inference, students do have some sound intuitions about data (Rubin et al., 2006; Bakker, 2004), which can be refined and moved towards reasoning that has inferential qualities (Rubin et al.).

According to Watson (2007), informal statistical inference represents a continuum of experience from the point when students start to pose questions about datasets to the point when they are about to meet formal inferential statistics. Along the way from informal to formal inference, a number of important ideas are added to the student package. Rubin et al. (2006) define informal inference as reasoning that involves consideration of the following related ideas:

(i) properties of aggregates rather than properties of individual cases,
(ii) sample size and its effect on the accuracy of population estimates or on process signals,
(iii) controlling for bias, and
(iv) tendency, distinguishing between claims that are always true and those that are often or sometimes true.

Ben-Zvi (2006) links informal inferential reasoning to argumentation and the need for data-based evidence. Deriving logical conclusions from data is accompanied by the need to provide persuasive arguments based on data analysis. Integration and cultivation of informal inference and informal argumentation seem to be essential in constructing students' statistical knowledge and reasoning in rich learning contexts (Ben-Zvi).

Zieffler, delMas, Garfield, and Gould (2007) view informal reasoning about statistical inference as the way in which students build connections between observed sample data and unknown or theoretical populations, and how they make arguments or use evidence to support these connections. Building such connections between sample and population lies at the heart of informal statistical inference (Johnston-Wilder, Ainley, & Pratt, 2007). Makar and Rubin (2007) regard informal inferential reasoning in statistics as a reasoned but informal process of creating or testing generalizations from data that extend beyond the data collected. They consider the following three principles to be essential to informal inference:

(i) Making generalizations (predictions, parameter estimates, conclusions) that extend "beyond the data,"

(ii) using data as evidence for these generalizations, and

(iii) using probabilistic language in describing the generalizations, including references to levels of certainty about the conclusions drawn.

Recent advances of technology provide schoolteachers and college instructors with new tools for adopting informal, data-driven approaches to statistical inference that can help lay the conceptual groundwork for formal inferential reasoning (Rubin et al., 2006). The appearance, in particular, of dynamic statistics learning environments (e.g., TinkerPlots and Fathom^{TM}), which are designed explicitly to facilitate the visualization of statistical concepts, provides enormous potential for making inferential reasoning accessible to students. These new technological tools provide a medium for the design of activities that integrate experiential and formal pieces of knowledge, allowing students to make direct connections between physical experience and its formal representations (Pratt, 1998; Paparistodemou, Noss, & Pratt, 2008). Students can experiment with statistical ideas, articulate their informal theories, use the theories to make conjectures, and then use the experimental results to test and modify these conjectures. Several researchers have, in recent years, been exploiting the affordances provided by these modern technologies for promoting learners' ability to reason and argue about data-based inferences, with very encouraging results (e.g., Bakker, 2004; Ben-Zvi, 2006; Meletiou-Mavrotheris, 2003; Rubin et al., 2006).

Some of the studies have demonstrated that even young children can develop powerful notions about inference when using appropriate data visualization tools. Ben-Zvi (2006) studied fifth-graders' learning processes in an exploratory interdisciplinary learning context that used the dynamic software TinkerPlots to scaffold and extend students' statistical reasoning. He described how the unique features of the software, in combination with a carefully designed learning trajectory based on growing samples heuristics, supported young learners in deriving persuasive arguments based on data analysis. Pratt (2000) found that 10-year-old students working with the *Chance-Maker* microworld were able to develop deep understanding of the ways in which empirical probability, theoretical probability, and sample size are related to the drawing of valid inferences. Stohl and Tarr (2002) described how two average-ability sixth-grade students were able, through using a variety of microworld tools, to formulate and evaluate inferences based on simulation data.

Although the emergence of studies specifically focusing on informal inferential reasoning has begun to shed some light on this important aspect of statistical reasoning, research on the topic is still at an embryonic stage. Since the investigation of informal inferential reasoning is a very recent endeavour, research "is only just grappling with understanding the conceptual building blocks for informal inferential reasoning as a pathway towards formal statistical inference," both in terms of the discipline and in terms of student cognition (Pfannkuch, 2006). In particular, there exists a gap in knowledge regarding young learners' informal notions of statistical inference. The current study contributes to bridging this gap by providing insights into ways in which instruction can facilitate young students' development of informal inferential reasoning through the provision of appropriate, technology-rich instructional settings that allow them to formulate and test conjectures regarding real datasets of interest to them.

3. PURPOSE/GOALS OF STUDY

The study reported in this article aimed at fostering third-grade students' informal notions of inference through adopting a hands-on, project-based approach to statistics using the dynamic statistics software TinkerPlots as an investigation tool and social activity focusing on making data-based arguments (Stohl & Tarr, 2002). This research is part of a larger, ongoing multifaceted program for the teaching and learning of early statistical reasoning in Cyprus, designed in response to the high level of interest in statistical reasoning and the need for further research on its development across grades and curricula. The conceptual "Framework for Teaching Statistics within the K-12 Mathematics Curriculum" (GAISE report, 2005), guided the program design. This framework focuses on building learners' conceptual understanding of the statistical process by emphasizing and revisiting, with increasing sophistication through the grade levels, a set of central statistical ideas. It uses a spiral approach to the statistics curriculum, so that instructional programs from pre-kindergarten through high school encourage students to gradually develop understanding of statistics as an investigative process that involves the following components:

(i) clarifying the problem at hand and formulating questions (hypotheses) that can be answered with data,

(ii) designing and employing a plan to collect appropriate data,

(iii) selecting appropriate graphical or numerical methods to analyze the data: summarizing the data, making conjectures, drawing conclusions, making generalizations, and

(iv) interpreting the results of the analysis and relating the interpretation to the original question.

Consistent with a theoretical perspective on statistics instruction at the school level, the study was focused on building young learners' informal inferential reasoning by enabling them to experience and develop the "big ideas" of statistics through the collection and exploration of real data of interest to them. The term informal inference is used here to describe the drawing of conclusions from data that is based mainly on looking at, comparing, and reasoning from distributions of data (Pfannkuch, 2006). Based on a case study of a group of 8-year-old students, the following questions regarding young learners' development of informal notions of inference were explored:

(i) How do young learners begin to reason about informal inference in a learning environment that adopted a hands-on, project-based approach to statistics?

(ii) How can the opportunities provided by a dynamic learning environment for formulating and justifying data-based inferences, be utilized in the early years of schooling to scaffold and extend students' informal statistical reasoning (Ben-Zvi, 2006)?

4. METHODOLOGY

4.1. CONTEXT AND PARTICIPANTS

The study took place during the 2006 Fall (Autumn) semester in a third-grade mathematics classroom in an urban primary school in Cyprus. Twenty-two students (about 8 years old) participated in this research. The students' reasoning about informal inference was studied through genuine statistical endeavours using the dynamical statistics environment TinkerPlots.

TinkerPlots is a recently developed dynamic data-visualization package intended primarily for elementary and middle grades. The software offers an easy-to-learn interface that encourages student activity. Using TinkerPlots, young learners can start exploring data without having knowledge of conventional types of graphs or of different data types. Through performing simple actions such as ordering data according to the values of a variable or sorting data into categories, children can develop a wide variety of both standard graphical displays (e.g., bar graphs, pie charts, scatterplots), but also unconventional data representations of their own invention (Ben-Zvi, 2000). They can progressively organize data to answer their questions. TinkerPlots aims at genuine data analysis with multivariate data sets from the start, by beginning with students' own ideas and working towards conventional statistical notions and graphs (Bakker, 2002).

Students participated in data-centered activities, in contexts familiar to them. After collecting real data about themselves, they worked in small groups to explore the data and to formulate and evaluate data-based inferences, using TinkerPlots as an investigative tool. These children had some previous experiences with graphing 'by hand', like collecting data from their class (e.g., the colours that their classmates like), making simple graphs (usually object graphs or bar graphs) by using grid paper, and drawing conclusions like finding the most popular colour amongst the children of their class and the number of children that liked this colour. They had no previous exposure to TinkerPlots other than going through a short tutorial which introduced them to the main features of the software. They became acquainted with TinkerPlots while analyzing data in this study.

The role of the researchers during the activities was that of *participant observers.* This stance indicates that as well as observing through participating in activities, the observer can ask the subjects to explain various aspects of what is going on (Burgess, 1984). The researchers were interacting with the students while they worked in mixed groups on the activities, and they were probing the children in order to better understand their thinking beyond their actions. The role of the researchers fell into probing interventions, experimental interventions, and technical interventions (cf. Pratt, 1998; Paparistodemou, 2004). Probing interventions aimed to make children's thinking transparent when it came to inferring the reasons that might lie behind their actions. Experimental interventions sought to make some change in the directions of the activity with possible implications for conceptual change. For example, whenever the opportunity arose, the researchers stepped in to encourage students to draw generalizations which extended beyond the data at hand. Technical interventions were made to give explanations about TinkerPlots software.

A case study design was employed in the study. It was judged that this research strategy was well suitable to exploring, discovering, and gaining insight into young children's perceptions, actions and interactions with the dynamic software TinkerPlots and with each other. The study was exploratory in nature, since very little is currently known about children's notions regarding informal inference. Thus, its purpose was not to prove or disprove hypotheses, but rather to generate descriptions based upon in-depth investigation of students' interactions with the technological tool and with each other, and of the impact these might have on their perceptions regarding informal inference. These descriptions, although of limited generalizability (Wegerif & Mercer, 1997), may be used to understand similar situations and can inform future research.

4.2. INSTRUMENTS/TASKS AND PROCEDURE

The main data source for the study's activities was a survey developed and administered by the children participating in the study, which investigated the health, nutritional and safety habits of students in their school. 'Nutrition, Health and Safety' was a school project theme for the entire month (November 2006). Students were introduced to this topic by learning about nutritional, health, and safety habits in a cross-curriculum environment. Different subjects in the school curriculum (e.g., science, art, music, language, and mathematics) emphasized different aspects of the school-project theme.

Students participating in this research first completed a personal diary about their nutritional, health, and safety habits. In completing the personal diary, children needed to have knowledge about what is a nutritional habit, a health habit, or a safety habit. This knowledge was gained through children's involvement with the other subjects in the school. After completing their personal diary, children then decided to compare their habits with those of their classmates. Finally, they started thinking about conducting a survey of the students of the school in order to present their results to a school fair at the end of the month. They decided that, in order to collect data from the first, second, and third-grade students of the school, they had to construct a more structured survey.

A 16-item questionnaire about gender, age, nutritional habits, health, and safety was prepared and administered by the students themselves. A total of 120 students completed the questionnaire. The data were then entered into a TinkerPlots database. Next, students employed the features of the dynamic statistics environment to explore and visualize the data, and to formulate and evaluate conjectures based on data (GAISE Report, 2005). They wrote a report and made a presentation of their findings to the whole school.

The study aimed at fostering, while at the same time also investigating, students' ability to collect and represent data and to propose and justify conclusions and predictions based on data (National Council of Teachers of Mathematics, 2000). Students were exposed to genuine data collection and analysis, computer-based experimentation, intensive use of visualizations, group work, and whole class discussions.

The research took place over a period of four weeks. There were 3-5 sessions per week, and each meeting lasted for about 40 minutes. During the study, the research team collected and analyzed a wealth of data to assess students' growth in understanding and reasoning about inference. Audio-recordings of class sessions, researchers' observations, videotaped interviews with selected students (the interviewing took place while students were working in groups analyzing their data), students' notebooks, and the students' final presentation to the whole school were used in the analysis.

The data collected during the course were first examined globally and brief notes were made to index them. The goal of this preliminary analysis was to identify representative parts indicative of students' approaches and strategies when performing specific statistical tasks. The selected occasions from videotapes were viewed several times and were transcribed. The transcribed data, along with other data collected in the study, were analyzed to determine the extent to which students had developed informal reasoning about statistical inference (Stohl & Tarr, 2002).

The study sought to identify and understand students' interactions with the dynamic statistics software and each other, and the ways in which these interactions influenced their inferential reasoning. Through empirical analysis of the data, inductively derived descriptions and explanations were obtained. These descriptions formed the study findings that are outlined in the next section.

The initial coding scheme of the transcripts was based on our interpretations of the existing literature review (e.g., GAISE Report as it is described in Section 3). The final coding system of the transcripts, which emerged after many sweeps through the data, was based on three general groups: (a) data-based argumentation, (b) data-based argumentation and generalization, and (c) data argumentation and chance. *Data based argumentation* refers to children's conclusions based on the data they had collected; *Data-based argumentation and generalization* refers to children's conclusions about their data and using the data to draw inferences about a larger population without engaging the idea of chance; and *Data argumentation and Chance* refers to children's conclusions about and using the data to draw inferences about an unknown population by engaging the idea of chance (e.g., using expressions such as 'more likely', 'might be', 'more possible to').

After analyzing the data using the particular coding system that we employed in our study (see also Paparistodemou & Meletiou, 2007), we came across Makar and Rubin's (2007) framework which refers to three key principles of informal inference—generalizations 'beyond the data', probabilistic language, and data as evidence. The authors used primary-school classroom episodes and excerpts of interviews with the teachers to illustrate the framework and reiterate the importance of embedding statistical learning within the context of statistical inquiry. Makar and Rubin's framework empowered our coding system of analysis for understanding young learners' informal inference.

5. RESULTS

Students decided to analyze their data using TinkerPlots, in order to present their findings at the school fair. The first researcher helped children to code their data on TinkerPlots data cards. Their data cards consisted of the following 16 attributes (see appendix for the questionnaire): gender, grade (1st year: Grade A, 2nd year: Grade B, 3rd year: Grade C), breakfast (what they eat for breakfast), snacks (what they eat between meals), lunch, dinner, the food they eat most often, eating sweets (how many sweets they eat every day), eating fruits (how many fruits they eat every day), exercising (if they play any sports), sleeping (what time they go to bed at night), teeth brushing (how many times they brush their teeth every day), using zebra crossing, running on the stairs (whether they run on school stairs), swinging with the chair (whether they swing on their chair in their classroom), playing with scissors (whether they play with scissors in class).

Children's interest on this task was high as they were very much involved with their school project. The following group of students, for example, gives a good justification as to why their class had decided to conduct an anonymous survey:

Researcher (R): What did you find interesting about the survey?
Chris: I remember that one boy was eating pizza for breakfast!
R: Was it something that he told you as a joke, or was it true?
David: I think that it was true.
R: Why do you think so?
David: Because we asked them not to write their name on the questionnaire, not to think that we are going to shout at them.
Danai: So, they did not write down their name in order not to tell any lies.

At this age, personal experience and interest play a key role in children's interactions with data. Personal interest played a key role in motivating children to get actively involved with the project, and to start reasoning about informal inference.

The snapshots presented in the following sections come from students' interactions with the software and with each other while analyzing the data. These snapshots show children's informal statistical inferential reasoning while trying to derive conclusions from data. The emphasis is placed on data-based argumentations, data-based argumentation and generalization and data argumentation and chance. On the graphs presented below, children used english characters for greek words, as there were not greek characters on the software.

5.1. DATA-BASED ARGUMENTATION

Data-based argumentation refers to children's conclusions based on the data they had collected. In the following snapshot, a group of three children is trying to analyze responses to the question "Do you play with scissors in class?" ("OXI"=NO, "NAI"=YES). They first create a bar graph showing that 113 out of the 120 children that completed the survey did not play with scissors in class (Figure 1).

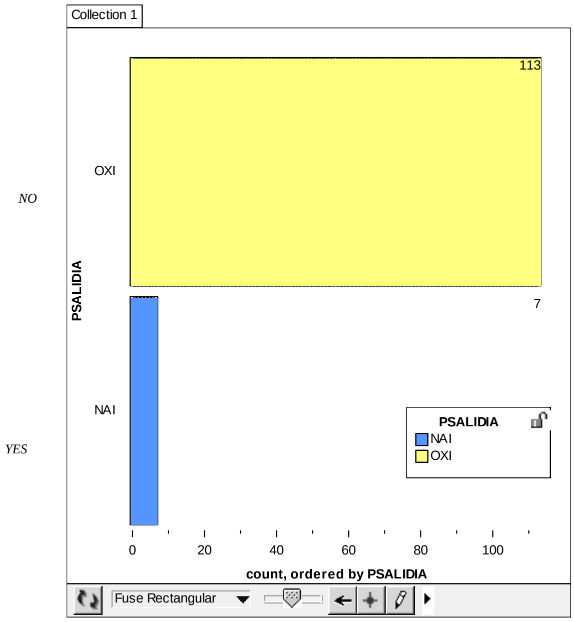


Figure 1. Answers to the question "Do you play with scissors ('PSALIDIA') in the class?"

R:	What can we see here?
Basil:	Most of the children are not playing with scissors.
Philip:	Playing with scissors is very dangerous… Some of the children are playing with scissors. Seven of them…
Basil:	Most of the children who are playing with scissors belong to Grade A.

The children's first reaction to this graph is to make the general conclusion that most of the children do not play with scissors. They relate this data-based conclusion with their personal experience that playing with scissors is dangerous. They also draw a conjecture based solely on their personal beliefs, that the majority of children playing with scissors belong to Grade A. The researcher intervenes here, prompting children and encouraging them to support this conjecture with data.

R: Do you have any evidence about this?
Basil: Yes … (he is trying to make a graph)
R: What are you trying to do?
Basil: To put the grade … (see Figure 2)
Basil: It is not only Grade A students who are playing with scissors, but also Grade B and Grade C students.
R: Of students playing with scissors, how many are in Grade A?
Basil: Four. [Actually in the graph there are only three children in Grade A playing with scissors.]
R: In grades B and C?
Basil: Two.
R: Is that a big difference?
Basil: No … it is not a big difference.
Mary: Most of the children in Grade A are not playing with scissors … the same for Grade B and Grade C.
Basil: But most children from the ones that are playing with scissors belong to Grade A.

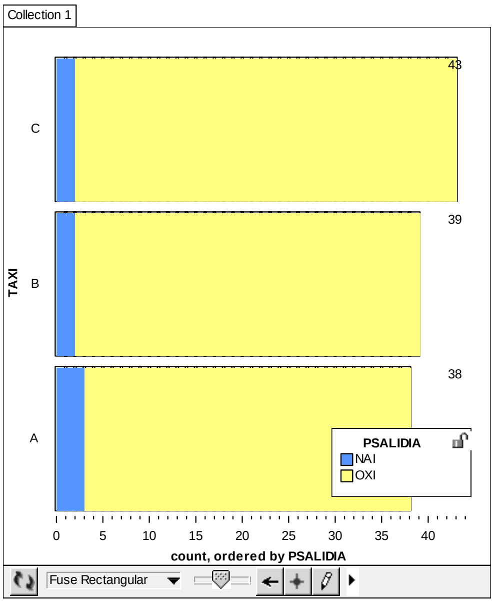


Figure 2. A fuse rectangular graph for grade ('TAXI') and playing with scissors ('PSALIDIA')

Mary compares Grade A with the other two Grades and, based on data, draws the conclusion that the majority of children in all three grades do not play with scissors. Basil attempts here to find evidence in the data to justify his personal belief that most of the children who play with scissors belong to Grade A. It is possible to distinguish here the juxtaposition of using personal experiences (like younger children do not know that scissors are dangerous) and the data as 'seen' in TinkerPlots. Interacting with TinkerPlots, Basil realizes that it is not only Grade A students who play with scissors. Although agreeing that differences in the number of children playing with scissors in different grades are small, he still emphasizes that his personal belief was correct.

5.2. DATA-BASED ARGUMENTATION AND GENERALIZATION

Data-based argumentation and generalization refers to children's conclusions about their data and using the data to draw inferences about a larger population without engaging the idea of chance. In the following excerpt, another group of students is trying to determine whether children in their school are exercising. Specifically, these students are analyzing responses to the survey question "Do you play any sports?" (see Figure 3).

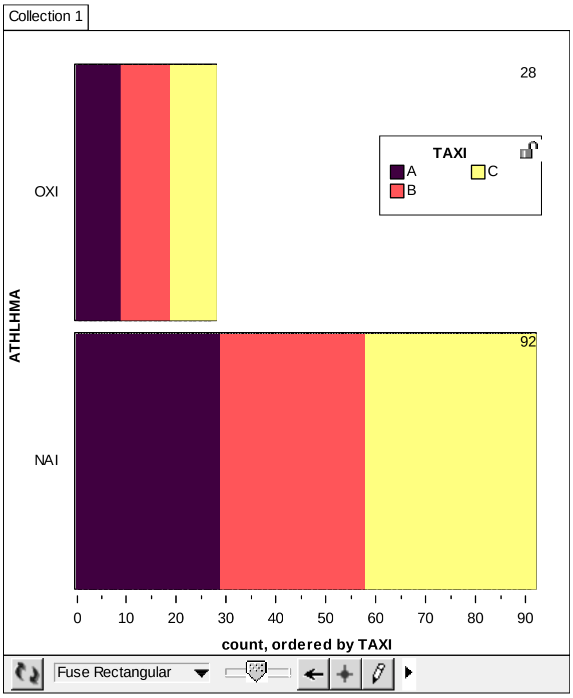


Figure 3. Exercising ('ATHLHMA' [sic]) in different grades ('TAXI'), ("OXI"=NO, "NAI"=YES)

R:	What can you see on the graph?
Melisa:	We can see how many children are exercising in each class.
R:	What can you conclude about exercising?
Bob:	We can see that most of the children are exercising.
R:	Who are these children?
Margaret:	The children of grades A, B and C.
R:	What can we say about grades?
Melisa:	Most of the children that are exercising belong to grade C.
R:	Ah! Can we conclude something else about grades?
Bob:	Yes! Children in grade C are exercising more than children in grade B and grade A.
R:	If we wanted now to include in our study the whole school, to include in our data also grades D, E, F what do you think we would conclude?
Bob:	The older you are in the school the more you exercise.

Melisa: That most of the children in the whole school are exercising, but also that some, not many, children are not exercising.
R: Who will be more?
Margaret: The children who are exercising.

The students in this snapshot draw conclusions about exercising in relation to grade level. Children provide here a data-based argumentation as to why they concluded that children in their school do exercise. In the children's words, the seeds of inferential reasoning are seen. The researcher decided to work with children further by asking the them to consider whether they thought that the data they had (and had already discussed) would be similar to what one would find in other classes (data they didn't have). Although the speculation about larger sets was occurring on the initiative of the researcher, it is interesting how the individual analysis of each grade led them to the conclusion that older children in their school are exercising more. Based on their data and from individual analysis, children drew conclusions for the distribution of the data set of a bigger population.

This group of children continues by comparing their own, individual responses to the survey data.

R: Do you remember what you had put down in your questionnaire?
Margaret: Yes … that I am exercising … so most of the children will be exercising. We need to exercise…
R: How many children are not exercising?
Bob: 28 and the ones who are exercising are 92.
R: Did you expect to have children that do not exercise?
Bob: I did expect it. Because the girls do not have a sport to do.
R: Do you think that it is only girls who do not exercise?
Bob: And some boys … I think that for a young child it is difficult to exercise.
R: Are the children who do not exercise only from grade A?
Bob: And from other grades … they might not know how to exercise …
Margaret: I noticed something! Grade B and Grade C might have the same number of children.
Melisa: We can see the numbers … in Grade A there are more children who are not exercising. [The difference between grades is not that big.]
Margaret: I also see the number of boys and girls in each class.

In this snapshot, peer-interaction seems conducive to deriving conclusions from data. The children here look at their graph axes in order to compare frequencies based on gender. In this snapshot it is possible to recognize once more the use of personal experience, which is used for justifying the data as 'seen' in TinkerPlots. The children bring their personal experience in making sense of data. Although the difference between the children who are exercising is not that big, Melisa uses the numbers on the graph to support her personal experience. Moreover, Margaret wants to explore their data further. She goes ahead to draw a new graph (Figure 4).

Margaret: I notice that in the 'No' answer there are more girls than boys.
Bob: The boys are fewer than the girls.
R: What do you mean in the 'No' answer?

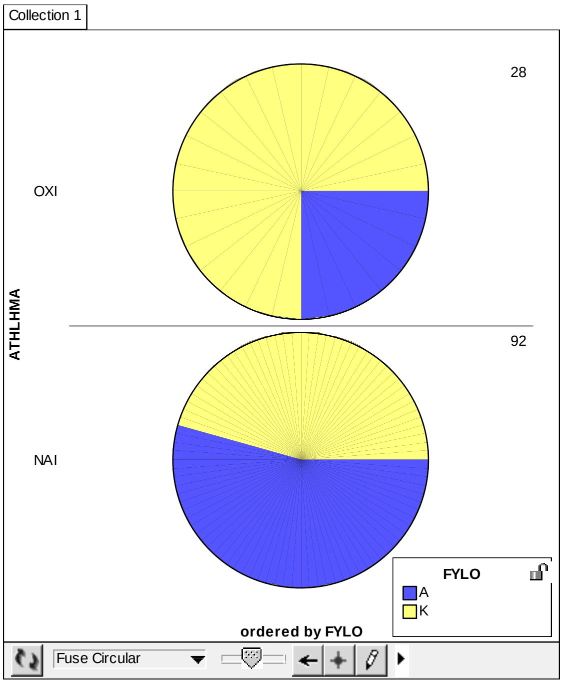


Figure 4. Exercising ('ATHLHMA') and gender ('FYLO'=GENDER: "A"=BOY, "K"=GIRL)

Margaret: That from the children that are not exercising, there are more girls than boys.
Melisa: Most of the children who are exercising are boys, while girls are not that many.
R: So …
Bob: Most of the children who are exercising are boys and most of the children in our survey are exercising.
R: If you go to another school what do you think you will notice?
Bob: The same thing …
R: What?
Melisa: Boys are exercising more than girls.
R: What about if I give the survey to all the schools in Cyprus?
Margaret: The children in Cyprus are exercising.
Bob: Not all of them …
Margaret: Half of them …
Bob: Not half of them. … Only one quarter of the children in Cyprus are not exercising.
R: Why?
Bob: Because we see here that most of the children in our school are exercising.
R: But, can we say the same thing for all the children of Cyprus?
Bob: Yes … we see it … most of the children would like to exercise. If not, they would get bored.
R: So, you are saying that only one quarter of the children in Cyprus is not exercising. From this quarter, what would you say about boys and girls?

Bob: Girls will be more.
R: Why?
Bob: Because they do not have a sport to do. Only dance and ballet . … Also, from the children who are not exercising, children in grade A are more, because they are not enthusiastic about sports yet.

The children go beyond their data and draw general conclusions about all Cypriot children when prompted by the researcher. Again, the speculation about larger sets occurred only on prompting of the researcher, but it is interesting that the children elaborate on the scenario and try to give a general percentage of the whole population of children who exercise in Cyprus. It is not surprising that at the beginning of this snapshot children make additive comparisons rather than proportional reasoning in comparing groups. But, when they talk about 'half' and use the pie charts, the children use intuitive ideas based on "part-whole" relationships seen in the graphs. The presence of the word "of" in the description is significant as it usually flags the part-whole understanding present (Watson, 2006). Moreover, they use their personal experience to give explanations for the results they observe in the data.

This group of children continues to interact with TinkerPlots in order to interpret their data (see Figure 5).

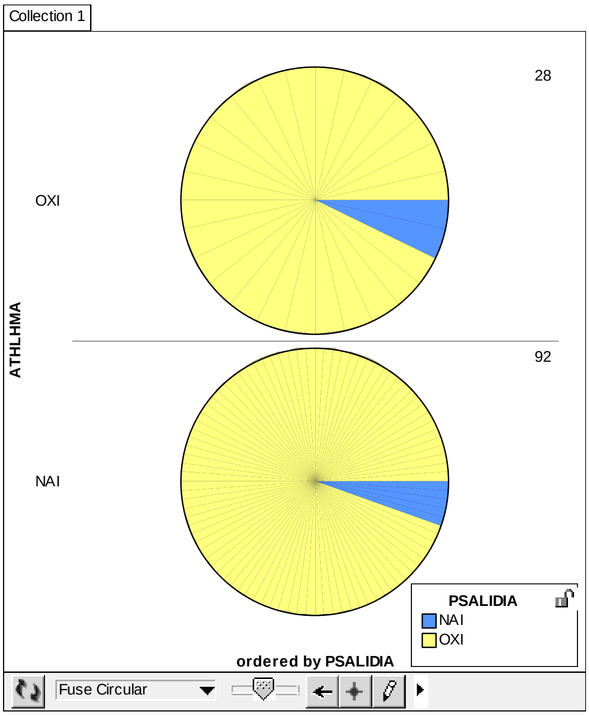


Figure 5. Exercising ('ATHLHMA') and playing with scissors ('PSALIDIA')

Melisa: Oh! I have done something else!
R: What have you done?
Bob: Ah! We can see here if the children who are exercising are playing with scissors.
R: Great! And what can you conclude?

Bob: Children who are exercising are not playing so much with scissors, but there are some children who are playing with them. But, of the children who are not exercising, there are more children who are playing with scissors.

R: Why?

Bob: Because we see it here. [He is showing on the graph.]

Melisa: They do not know that it is not good for them to play with scissors.

Bob: Ah! Maybe with scissors they cut themselves and they cannot exercise!

R: If we give this survey to the whole school [Grades A-F] will we notice the same thing?

Bob: No!

R: Why?

Bob: Because the children will be older and they will know that they should not do these things.

R: If we go to another school that has only the lower grades? [Grades A-C]

Melisa: Yes! I think grade A children will be playing more with scissors…

Bob: I would like to see a graph about grades and scissors.

R: Ah! Can you make it?

Bob is trying to make an argument about the relation between playing with scissors and exercising. His initial conclusion is based on the graphs, although at the end he and Melisa are creating a fictitious scenario in order to justify their data. Creating this scenario is the reason the children over-generalize their data, but on the other hand it is also the reason to drive Bob's curiosity about finding out what the data show regarding the relation between playing with scissors and grade level. Bob goes ahead and draws the pie-charts in Figure 6.

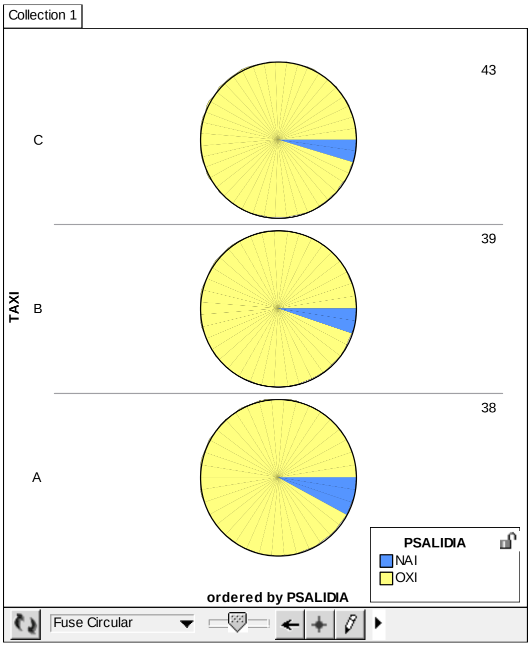


Figure 6. A pie-chart graph for grade ('TAXI') and playing with scissors ('PSALIDIA')

Margaret: I can see that Grade A children are playing more with scissors.
Bob: I prefer to see it on a rectangular graph. [Meaning a horizontal bar graph.] It shows it more clearly. [See Figure 2.]
R: What can you notice now?
Bob: That the children in Grades B and C who are playing with scissors are fewer than the children in Grade A. In Grade A, children are younger and play more with scissors. [He is pointing to the graph.]
R: If we ask all the children in our school?
Margaret: Most of them will not play with scissors.
R: What can we say about all children in Cyprus?
Bob: Again, one quarter…. No, less than one quarter are playing with scissors.
R: Where can we find this?
Bob: See the numbers … 4 and 3 and 3 from each class. [Actually it is 3, 2, 2 on the graph.] … not a big difference. …
R: So how many?
Bob: 10 from 120… so…
Margaret: 1 out of 12 is playing with scissors. Grades B and C are having the same number…

It is interesting here how Bob is using the different graphical representations of data provided by TinkerPlots. He finds it easier to compare grades in a rectangular representation than in a pie-chart, and he makes the argumentation that children in Grade A play more with scissors than children in other grades. When asked by the researcher to generalize their findings to all Cypriot children, the children attempt to make numerical arguments about the population of children in Cyprus. Again, an intuition based on 'part-whole' relationships is seen by the students in the graphs. Margaret uses the phrase '1 out of 12' to indicate the proportion of the children who are playing with scissors. The particular phrase has a direct relationship to the part-whole concept that is a feature of many topics in the mathematics curriculum.

5.3. DATA ARGUMENTATION AND CHANCE

Data argumentation and Chance refers to children's conclusions about the data and using the data to draw inferences about an unknown population by engaging the idea of chance. The next group of children analyze responses to the question "How many times do you brush your teeth every day?"

R: What can we say about teeth brushing in different grades?

Students draw the graphs displayed in Figure 7.

Natalie: Only a few children do not brush their teeth.
Demis: They do not brush their teeth every day.
Natalie: Most of them are in Grade B.
R: How do you know that?
Natalie: From the numbers.
Niki: … and the colours …

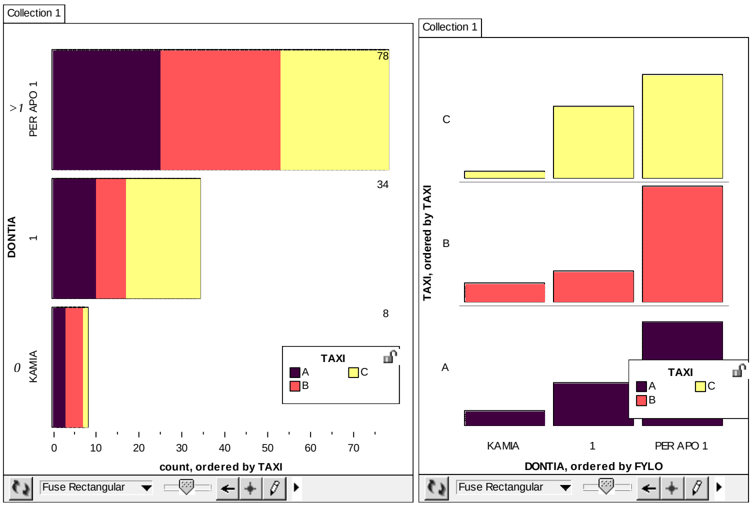


Figure 7. How often children brush their teeth ('DONTIA': "KAMIA"=None, "PER APO 1"=more than once) in different grades ('TAXI')

Demis: Most of the children brush their teeth … and in Grade B children do not brush them.
Natalie: In Grade C, there are a lot of children who brush their teeth only once a day.
R: What else?
Demis: Most of the children brush their teeth more than one time …
R: What else?
Natalie: The gender of the children … [She draws the graph in Figure 8.]

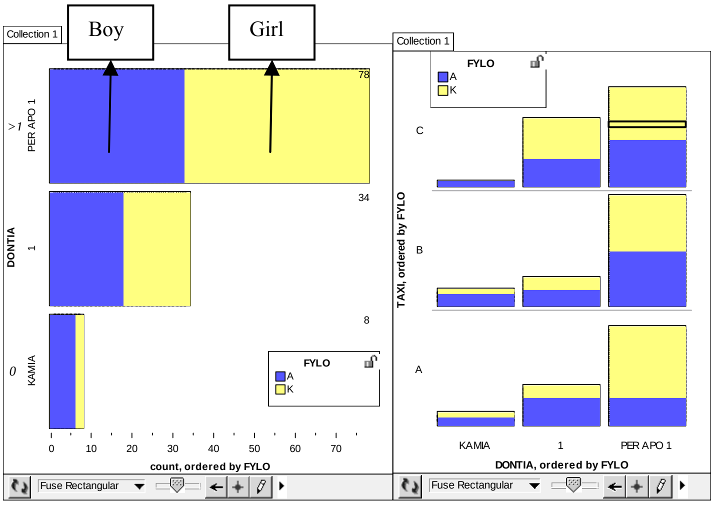


Figure 8. How often children brush their teeth based on gender ("FYLO"=GENDER: "A"=BOY, "K"=GIRL)

Natalie: Girls are brushing their teeth more than once. Boys are not brushing their teeth at all.

R: So, if you see a boy at break time, will you say that it is more likely that he brushes his teeth more than once or that he does not brush them at all?
Natalie: More than once.
R: Why?
(There is silence.)
R: If I tell you that I saw a child at break time that does not brush his/her teeth. Will you guess that is a boy or a girl?
Demis: A boy…
R: Why?
Natalie: We *see* it on the graph.
Demis: It might be a girl as well …
Natalie: But, most of the boys do not brush their teeth.
Demis: Yes. But, it might be a girl as well.
R: Is it more possible to be a girl or a boy?
Demis: More possible to be a boy!

Children in the above group describe the results for this question by looking at graphical representations of the data, but also try to connect their graphs with chance with prompting from the researcher. It is interesting that Natalie *reads from the graph* that a better guess for a child that does not brush his/her teeth is for it to be a boy. It is conjectured here that children in this group go beyond the data they have at hand, an important element for informal statistical inference. On the other hand, statements such as "But, most of the boys do not brush their teeth" lead to over-generalizations that make it difficult to draw any consistent conclusions.

The following group of children also tries to go beyond their data. They analyze students' responses to the survey question "How many sweets do you eat every day?" (see Figure 9).

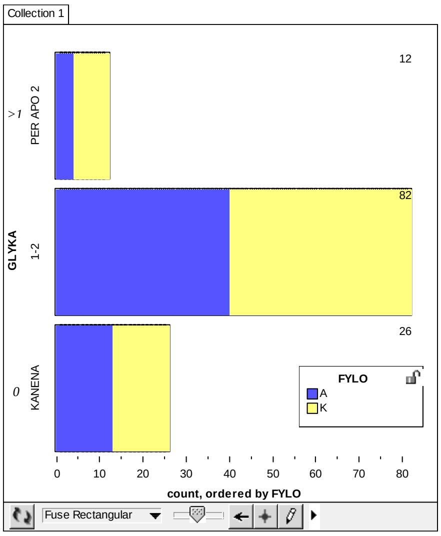


Figure 9. The number of sweets ('GLYKA':'KANENA'=None, 'PER APO 2'=More than 2) boys and girls eat ('A'= Boy, 'K'=Girl)

R: Can you make a graph and tell us your conclusions?
Stalo: For sweets … Girls are eating more sweets.

Stalo is making a general statement about the results.

…

R: If I find a child and I tell you that he/she eats more than two sweets, would you say it is a boy or a girl?

Stalo: A girl … Because this graph shows that the girls who eat more than two sweets are more than the boys.

R: If the child doesn’t eat any sweets?

Andreas: I think it is more likely for the child to be a boy, because fewer boys than girls eat more than two sweets, and it is more likely for boys not to eat any sweets.

R: Why?

Andreas: Because girls eat more than two. It is more likely for the child that does not eat any sweets to be a boy.

Stalo: I also say that the child is a boy; because boys do not eat many sweets … we eat more sweets.

In this group of students, a mixture of personal experience and observation of real data is seen. Again, there is a relation between data and chance and it can again be concluded that this group of children also go beyond their data.

6. DISCUSSION

Reflecting the recent shift in statistics education research from a focus on specific skills and procedures towards a greater focus on statistical reasoning and thinking embedded in the process of a statistical investigation (Makar & Rubin, 2007; GAISE Report, 2005), the current study was designed to investigate ways in which the foundations of inferential reasoning can be laid at a very young age. More specifically, the following two research questions regarding young learners’ development of informal notions of inference were explored. How do young learners begin to reason about informal inference in a learning environment that adopts a hands-on, project-based approach to statistics? How can the opportunities provided by a dynamic learning environment for formulating and justifying data-based inferences be utilized in the early years of schooling to scaffold and extend students’ informal statistical reasoning?

The study findings do support the view that statistics instruction can promote the development of learners’ inferential reasoning at an early age, through an informal data-based approach, rooted in the statistical investigation cycle (Wild & Pfannkuch, 1999). The design of the study, guided by the GAISE Report (2005), proved helpful in building children’s informal inferential reasoning. The 8-year-old students in the study experienced statistics as an investigative process. They formulated questions of interest to them, collected data to answer to these questions, analyzed and interpreted the data, linking their data-based conclusions and inferences back to the questions under investigation.

The young children in our study expressed statistical informal inference in three distinct ways: (a) data-based argumentation, (b) data-based argumentation and generalization, and (c) data argumentation and chance. Children drew their conclusions based on the data they had collected, using the data to draw inferences about a larger population without engaging the idea of chance, and using the data to draw inferences about an unknown population by articulating uncertainty (e.g., using expressions such as ‘more likely’, ‘might be’, ‘more possible to’). It seems that informal inference provides new opportunities to introduce powerful statistical concepts early in the school curriculum. This study shows that statistics can be used as a tool for gaining insights into understanding problems, rather than only as a collection of graphs, calculations, and procedures (Sorto, 2006).

Findings from this study show that young learners begin to reason about informal inference when their interest in the task is high. The children in this study were very much involved with their school project and the conclusions drawn from the data were important for them in order to understand what was happening at their school. At this age, personal experience and interest play a key role in children's interactions with data. Personal interest is important for children's involvement in reasoning about informal inference. Moreover, the study is an example of an approach to improving students' use of statistical reasoning and thinking by embedding statistical concepts within a purposeful statistical investigation that brings the context to the forefront. It is not just making a conclusion *about data* that provides the conceptual muscle to draw inferences, but a conclusion *about the situation* that the data are meant to represent or signify (Makar & Rubin, 2007). The focus is on understanding the situation (Makar & Confrey, 2007), rather than examining decontextualized data. Perhaps a focus on an interesting problem and engaging context may influence students' inclination to look beyond the data they have. Making a conclusion *about the situation* suggests that students need a particular level of complexity to engage with in order to consider possible avenues to connect the data with the context.

Moreover, the researchers in the study moved between referring to the data children had in hand and a larger population beyond the data. With prompting from the researchers, the children tried to draw conclusions "beyond the data," although sometimes they tended to over-generalize (for example, statements like "most of the boys do not brush their teeth"). Several times, when children tended to over-generalize the data, they tried to justify their 'scenario' by interpreting with the software. A result of that was to interpret with their data, make different graphical representations, and draw conclusions beyond the data (e.g., Bob's case). The students in our study used the dynamic statistics software TinkerPlots as an investigation tool. The presence of the dynamic software facilitated students' interest in the statistical investigation; it gave them the opportunity to explore data and draw data-based arguments and inferences in ways that would not have been possible for them without the software (Hammerman & Rubin, 2003).

Attributes of TinkerPlots like the ability to operate quickly and accurately, to dynamically link multiple representations, to provide immediate feedback, and to transform an entire representation into a manipulable object enhanced students' flexibility in using representations and provided the means for them to focus on statistical conceptual understanding. The visualization of the data helped children to express intuitive ideas about proportional reasoning, a fundamental topic in the school mathematics curriculum. The genuine endeavors of the young learners with multivariate data using TinkerPlots as an investigation tool helped them begin to develop their informal inferential reasoning. Furthermore, the software's design allows even young students to use what they already know to search for and detect group differences and trends. By using features such as differences in icon size, colour (e.g., the user can highlight information by the value of an attribute), students can detect subtle relationships in multivariate data in powerful and intuitive ways (Bakker, 2002). Although that the children in the study were young, most of the time they tried to find relationships between two variables in the data in order to draw their conclusions. For example, in Figure 5 they tried to find the relationship between exercising and playing with scissors.

The qualitative methodology employed in this case study, the small scale of the study, and its limited geographical nature, mean that generalizations to cases that are not very similar should be done cautiously. However, the study findings do suggest that the adoption of a hands-on, project-based approach to statistics does have the potential to enhance statistics instruction by making inferential reasoning accessible to young learners. Moreover, there are strong indications in the study to support our belief that utilization of the affordances provided by a dynamic statistics software such as TinkerPlots, can indeed scaffold and extend children's informal statistical reasoning (Ben-Zvi, 2006) by encouraging them to build, refine, and reorganize their intuitive understandings about statistics.

7. INSTRUCTIONAL AND RESEARCH IMPLICATONS

The expanding use of data for prediction and decision-making in almost all domains of life makes it a priority for mathematics instruction to help all students develop their inferential statistical reasoning. As the current study and several other studies have illustrated (e.g., Ben-Zvi, 2006; Makar & Rubin, 2007), when given the chance to participate in appropriate instructional settings that support the development of informal inferential reasoning, even very young children can develop intuitions about fundamental statistical concepts related to statistical inference.

In order to promote the development of early inferential reasoning, statistics instruction should adopt an informal, data-driven approach. It should encourage statistical inquiry rather than teaching methods and procedures in isolation. The emphasis should be on the statistical investigation process. The teaching of the different statistical tools should be achieved through putting students in a variety of authentic, purposeful contexts where they need these tools to make sense of the situation. Instruction should focus on helping learners understand how one could use these tools in making comparisons, predictions, and generalizations (Rubin, 2005). Through exploration and experimentation with authentic data, children can begin to develop the ability to provide persuasive data-based arguments, as well as generalizations which extend beyond their collected data.

Technology has an important role to play in promoting the development of informal inferential reasoning in the early grades. Innovative educational software such as TinkerPlots allows children to explore informal inferential ideas in contexts that are both rich and meaningful to them. They provide young learners with tools they can use to construct their own conceptual understanding of statistical concepts. Use of such software, in combination with appropriate curricula and instructional settings, can help students develop a strong conceptual base on which to build a more formal study of inferential statistics later.

More research is needed on designing instruction and on building teachers' and students' concepts and reasoning about informal inference. The challenge for future research should be to develop teachers' and students' sampling conceptions in terms of learning to reason about populations from samples using informal inference. The focus should shift from examining how students learn and use the different statistical tools and procedures, to how they understand the statistical investigation cycle as a process of making data-based inferences (Makar & Rubin, 2007). The study shows that students should come to appreciate the real reasons for which statistical tools and procedures have been developed—to help humans understand underlying phenomena and make informed decisions.

ACKNOWLEDGEMENTS

The authors wish to thank all the children that participated in this research. They would also like to thank all the participants of SRTL-5 conference for their very generous and constructive comments on this work. Moreover, they would like to thank Jane Watson for her valuable feedback on the manuscript and the reviewers of *SERJ* for their very detailed reviews.

REFERENCES

Bakker, A. (2002). Route-type and landscape-type software for learning statistical data analysis. In B. Phillips (Ed.), *Developing a statistically literate society*: *Proceedings of the Sixth International Conference on Teaching Statistics,* Cape Town, South Africa. [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/1/7f1_bakk.pdf]

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*(2), 64-83.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(2)_Bakker.pdf]

Ben-Zvi, D. (2000). Toward understanding the role of technological tools in statistical learning. M*athematical Thinking and Learning*, *2*, 127-155.

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D1_BENZ.pdf]

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics*, *45*(1), 35-65.

Burgess, R. (1984). *In the Field.* New York: Routledge.

GAISE Report (2005). *Guidelines for assessment and instruction in statistics education: A Pre-K-12 Curriculum Framework.* Alexandria, VA: The American Statistical Association. [Online: http://www.amstat.org/education/gaise]

Garfield, J., & Ahlgren, A. (1998). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*, 44-63.

Gordon, F. S., & Gordon, S. P. (1992). Sampling + Simulation = Statistical Understanding. In F. S. Gordon (Ed.), *Statistics for the twenty-first century* (pp. 207-216). Washington, DC: The Mathematical Association of America.

Green, D. G. (1982). A survey of probability concepts in 3000 students aged 11-16. In D. V. Grey (Ed.), *Proceedings of the First International Conference on Teaching Statistics* (pp. 766-783). London: Statistics Teaching Trust.

Hammerman, J., & Rubin, A. (2003). Reasoning in the presence of variability. In C. Lee (Ed.), *Reasoning about Variability: A Collection of Current Research Studies* [CD-ROM]. Dordrecht, The Netherlands: Kluwer Academic Publisher.

Johnston-Wilder, P., Ainley, J., & Pratt, D. (2007, August). *Thinking-in-change about informal inference*. Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5), University of Warwick, UK.

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: Seymour.

Konold, C., & Miller, C. D. (2005). *TinkerPlots: Dynamic Data Explorations*. Emeryville, CA: Key Curriculum Press.

Makar, K., & Confrey, J. (2007). Moving the context of modeling to the forefront. In W. Blum, P. Galbraith, H-W. Henn, & M. Niss (Eds.), *Modelling and applications in mathematics education*. New York: Springer.

Makar, K., & Rubin, A. (2007, August). *Beyond the bar graph: Teaching informal statistical inference in primary school*. Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5), University of Warwick, UK.

Meletiou-Mavrotheris, M. (2003). Technological tools in the introductory statistics classroom: effects on student understanding of inferential statistics. *International Journal of Computers for Mathematical Learning, 8*(3), 265-297.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and Standards for School Mathematics*. Reston, VA: Author.

Paparistodemou, E. (2004). *Children's Expressions of Randomness: Constructing Probabilistic Ideas in an Open Computer Game*. Unpublished doctoral dissertation, Institute of Education, University of London.

Paparistodemou, E., & Meletiou, M. (2007, August). Enhancing reasoning about statistical inference in 8-year-old students. Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5), University of Warwick, UK.

Paparistodemou, E., Noss, R., & Pratt, D. (2008). The Interplay Between Fairness and Randomness in a Spatial Computer Game. *International Journal of Computers for Mathematical Learning, 13*(2), 89-110.

Pfannkuch, M. (2006). Informal inferential reasoning. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/6A2_PFAN.pdf]

Pratt, D. (1998). *The Construction of Meanings In and For a Stochastic Domain of Abstraction*. Unpublished doctoral dissertation, Institute of Education, University of London.

Pratt, D. (2000). Making sense of the total of two dice. *Journal of Research in Mathematics Education, 31*, 602-625.

Rubin, A., Bruce, B., & Tenney, Y. (1990, April). *Learning about sampling: Trouble at the core of statistics*. Paper Presented at the Annual Meeting of the American Educational Research Association, Boston.

Rubin, A. (2005). Math that matters. In K. Mayer (Ed.), *Threshold* (pp. 22-31). Cambridge: TERC.

Rubin, A., Hammerman, J., & Konold, C. (2006). Exploring informal inference with interactive visualization software. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D3_RUBI.pdf]

Watson, J. M. (2006). *Statistical literacy at school* New Jersey: LEA.

Watson, J. (2007, August). Facilitating beginning inference with TinkerPlots for novice grade 7 students. Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5), University of Warwick, UK.

Wegerif, R., & Mercer, N. (1997). Using computer-based text analysis to integrate qualitative and quantitative methods in research on collaborative learning. *Language and Education*, *11*(4), 271–86.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review*, *67*(3), 223-265.

Shaughnessy J. M., Ciancetta M., Best K., & Canada D. (2004, April). Students' attention to variability when comparing distributions. Paper presented at the 82nd Annual Meeting of the National Council of Teachers of Mathematics, Philadelphia, PA.

Sorto, M. A. (2006). Identifying content knowledge for teaching statistics. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/C130.pdf]

Stohl, H., & Tarr, J. E. (2002). Using multi-representational computer tools to make sense of inference. In D. Mewborn (Ed.), *Proceedings of the Twenty-fourth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Athens, GA: Columbus, OH: ERIC Clearinghouse for Science Mathematics and Environmental Education.

Zieffler, A., delMas, R., Garfield, J., & Gould, R. (2007, August). Studying the development of college students' informal reasoning about statistical inference. Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5), University of Warwick, UK.

EFI PAPARISTODEMOU
5 Ellados Street
2003 Nicosia
Cyprus

APPENDIX: THE QUESTIONNAIRE

We are the children of C_2 class and we are doing a study that concerns nutrition, health and safety of the children in our school. We would like to ask you to complete the following questionnaire. Do not write your name anywhere. Thank you! ☻

1. Gender: Boy □ Girl □

2. Grade: A □ B □ C □

3. Nutrition:

3.1. What did you eat yesterday?

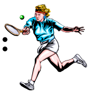
Breakfast: ....
Lunch: ....
Dinner: ....
Snacks: ....

3.2. Which food do you eat most often? …………………………

3.3. How many sweets do you eat every day? None □ 1-2 □ more than 2 □

3.4. How many fruits do you eat every day? None □ 1-2 □ 3-4 □ more than 4 □

4. Health:



4.1. Do you play any sports? Yes □ No □

4.2. What time do you go to bed at night? Before 9pm □ 9pm □ After 9pm □

4.3. How many times do you brush your teeth every day? None □ 1 □ More than 1 □

5. Safety at school:

5.1. Do you always use zebra crossing? Yes □ No □

5.2. Do you run on the school stairs? Yes □ No □

5.3. Do you swing on your chair in the classroom? Yes □ No □

5.4. Do you play with scissors in the classroom? Yes □ No □

LOCAL AND GLOBAL THINKING IN STATISTICAL INFERENCE

DAVE PRATT
Institute of Education, University of London
d.pratt@ioe.ac.uk

PETER JOHNSTON-WILDER
Institute of Education, University of Warwick
p.j.johnston-wilder@warwick.ac.uk

JANET AINLEY
School of Education, University of Leicester
janet.ainley@le.ac.uk

JOHN MASON
Centre for Mathematics Education, Open University
j.h.mason@open.ac.uk

ABSTRACT

In this reflective paper, we explore students' local and global thinking about informal statistical inference through our observations of 10- to 11-year-olds, challenged to infer the unknown configuration of a virtual die, but able to use the die to generate as much data as they felt necessary. We report how they tended to focus on local changes in the frequency or relative frequency as the sample size grew larger. They generally failed to recognise that larger samples provided stability in the aggregated proportions, not apparent when the data were viewed from a local perspective. We draw on Mason's theory of the Structure of Attention to illuminate our observations, and attempt to reconcile differing notions of local and global thinking.

Keywords: *Statistics education research, Task design, Informal statistical inference, Sample size, Local and global meanings or perspectives, Structure of Attention*

1. WHAT IS INFERENCE AND IN WHAT SENSE IS IT INFORMAL?

Statistical inference is typically introduced as a formal topic in the curriculum at around age 16 or older. Inferential analysis is typically taught as a tool for judging the source of variation in data. Students' lack of comprehension has been widely reported and in response there has been a recent research effort to understand how better to approach the topic from a pedagogic perspective. One response has been Exploratory Data Analysis, in which students attempt to infer informally about underlying trends in data without explicit reference to probability. These approaches make even more urgent a deeper understanding of how young students might make sense of this inferential process.

Statistics Education Research Journal, 7(2), 107-129, http://www.stat.auckland.ac.nz/serj

In this paper, we reflect on recent small-scale experiments to clarify understanding of young students' activity when engaged in an inference related task designed to trigger intuitions, assumptions and conceptual thinking. Our aim is to elaborate the conceptual struggle that needs to take place for young students to engage in inferential reasoning. In so doing, we acknowledge a constructivist stance in which we search for naïve conceptions (as opposed to misconceptions, a distinction delineated by Smith, diSessa, & Rochelle, 1993) that might serve as resources for further development. We begin by clarifying our perspective on what we see as the conceptual roots of statistical inference because it is in those roots that we may find informal inference.

Makar and Rubin (2007) give a working definition that we find useful: "We consider informal inferential reasoning of statistics in broad terms to be the process of making probabilistic generalizations from (evidenced with) data that extend beyond the data collected." For us, this definition describes, without misrepresentation, statistical inference per se. By focussing on the conceptual basis for inference, we can begin to imagine younger students engaging in such activity without necessarily conducting formal statistical hypothesis tests or building carefully defined confidence intervals, as we would see in a classical statistics course. Our focus is primarily on students around the age of 10 or 11 years, well before any formal teaching of inference. It is therefore likely that intuitions will not yet have been formalised through schooling and that student thinking can be thought of, in that sense, as informal.

However, we wish to probe a little further into what Makar and Rubin might mean by "making probabilistic generalisations from data." Immediately, we recognise a direction to the statement – from data to probabilistic generalisation. Inference is concerned with identifying patterns in the form of trends or statistical parameters in the "underlying" population. There has been recent research interest in how children might make, or be encouraged to make, inferences such as these by studying how children think informally about populations, given samples of data (Ben Zvi, 2006; Pfannkuch, 2006). Typically, the activity has been focussed on using sampled data to make statements about a finite population from which the data were drawn. However, other possibilities for informal inference exist. Our interest focuses on situations where the population cannot be described in terms of the total finite dataset but instead is most adequately described through a probability distribution. The data may be used to draw inferences about aspects of an infinite population or process. Indeed, our focus in this paper will largely be on inferences about the probability distribution associated with outcomes from a die. Here again it is important to distinguish between an expert sophisticated perspective on what was happening and the likely relatively naïve perspective of a 10-year-old. Although we might see the activity as being rooted in making inferences about a probability distribution, it is reasonable to think that 10-year-olds saw the game they were playing as trying to guess what the die looked like.

We also note that in some reported studies on informal inference the activity used may be open to ambiguous interpretation. During SRTL-4 in Auckland, New Zealand, the theme for discussion was students' thinking about distribution, rather than inference. However, several presentations showed students working informally with data. There was a realisation during discussion that the focus of the students may at times have been on the dataset as if it were the whole population, whereas the teacher's attention may have been on the underlying population from which that dataset was only a sample. We referred to these two situations as Game 1, where the dataset was all there was, and Game 2, where the dataset was merely a sample. This distinction is critical in considering how students think about inference. Game 1 allows no room for inference (*what you see is what you get*) whereas Game 2 demands an ability to make inferences about the population on the basis of the data available (*what you see* might be *what you get*). We designed a task involving making inferences about the probability distribution of a die, in which, even from the students' perspective of simply guessing the configuration of the die, the focus is firmly in Game 2. The results of throwing the die are a sample arising from a "hidden" (or "underlying" in the usual statistical parlance) generator (or "population").

Let us now return again to Makar and Rubin's definition, and consider the thinking involved in 'making probabilistic generalizations from data'. Prodromou (2007), and Prodromou and Pratt (2006) have shown that 15-year-old students, using a Basketball microworld, were able to make connections between data and what was called the modelling distribution. (We can think of the modelling distribution as a probability distribution used to model any particular phenomenon. Prodromou distinguished two perspectives on distribution, the modelling perspective and the data-centric perspective, which attends to the sample of data.) She has proposed the mechanisms that act in the explanations expressed by learners as quasi-causal agents connecting the modelling and the data-centric perspectives. These agents are not causal in the sense of direct cause and effect. Rather they are invented substitutes, which for learners play the role of causal agents given the lack of any explicit determining agent.

In the 'inferential direction', from a data-centric to a modelling perspective, some of Prodromou's students saw the modelling distribution as a target, towards which the data-centric distribution was aiming. There is a sense of variation in the data, out of which the population is an emergent phenomenon, like a trend (Prodromou, 2008). In the absence of any explicit cause for how this happens, emergence itself is seen as a somewhat mysterious causal-like agent that enables the modelling distribution to emerge out of the data-centric distribution.

In the opposite direction, from a modelling to a data-centric perspective, there is a sense of intention. In Prodromou's research, the intention is attributed by the students variously to the human modeller, the characters in the software or the tools within the software that trigger the random generation of data from the modelling distribution. Software tools allowed students to explain the generation of the data-centric distribution as 'caused' by the actions of agents on behalf of the modelling distribution. We would recognise such explanations as the situated roots of a view of probability theory as a causal-like agent that enables the modelling distribution to generate the data-centric distribution. Indeed, in classical statistics, data are seen as generated through a mixture of signal and noise from a modelling distribution.

Although we would regard making such connections between data and population as lying at the heart of informal statistical inference, neither probability nor emergent phenomena are trivial areas of mathematical modelling, and so, when we observe children, we are likely to see either (i) naïve understandings, which serve to make sense of the world in the absence of such theory, or (ii) meanings, which appear to us to be rooted in a situational way to these abstracted theories. Maker and Rubin (2007) refer to the importance in informal inference of aggregate thinking, sample size, controlling for bias, and tendency. Our focus is firmly related to aggregate thinking and how students might perceive the effect of sample size on the inferential process. In particular, we have become interested in the shifting of attention between what is happening in the here-and-now, the immediate, and what is happening in an aggregated sense over the longer-term.

2. THE LOCAL AND THE GLOBAL

Our interest in informal inference emerges out of aspects of previous work. Pratt and Noss (2002) reported that 10- to 11-year-olds were able to articulate expert-like views about short-term randomness, through meanings that were immediately accessible. These meanings were described in that study as *local*, in the sense that such resources focus on trial by trial variation. According to Pratt and Noss, local meanings refer to unpredictability, irregularity and uncontrollability.

In contrast, these students did not easily express meanings for long-term randomness. Using a design research approach (Cobb et al., 2003), Pratt built a computer-based domain of stochastic abstraction called *ChanceMaker*. This microworld provided 'gadgets', simulations of everyday random generators such as coins, spinners and dice, whose behaviour was controlled through a 'workings box', which is an unconventional representation of the distribution. Figure 1 shows the default workings box for the dice gadget. The software allowed children to explore the gadgets using tools to display results in different ways (pie chart, pictograms, and lists), and to examine and edit the contents of the workings box, which controlled the behaviour of the gadget (see Figure 1). Challenged to identify which gadgets might not be working properly, the children began to use the tools provided to decide how to mend the gadgets.

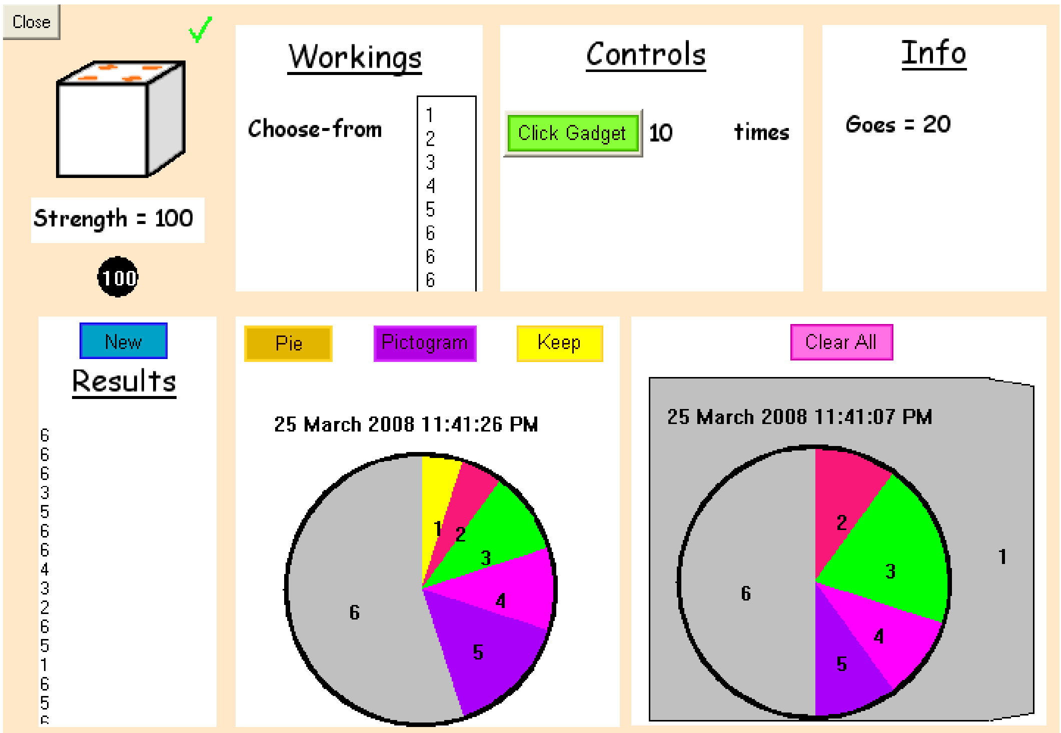


Figure 1. The ChanceMaker dice gadget can be opened up to reveal a "broken" workings box. Here the student has created a pie chart from 10 throws, kept that picture, and then created a second pie chart from 20 throws.

Gradually, the children began to articulate meanings for long term randomness, which focussed on an aggregated overall view of the stochastic, such as *the proportion of outcomes was predictable* (probability), *the proportion of results stabilised when more* *trials were executed* (large numbers), and *the observer is able to exert control over these proportions through manipulation of the possibility space* (distribution).

Pratt (2000) noted an interesting correspondence between local and global meanings:

> "…local resources tend to be inverted in relation to their global counterparts. Thus, unpredictability as a local resource is inverted in comparison to the global resource of predictability (in a proportional sense). Similarly, control cannot be exerted locally whereas there is a global resource for control through manipulation of the distribution." (p. 609)

In Pratt's study, the students were observing patterns in the data via the different representations. By interacting with the workings box, they were in effect creating a probability distribution to generate data. Their task became one of understanding the nature of the control exerted by the workings box over the variability and the structure of the data. It is clear from this work that it was far from trivial for students of this age to identify the global structures that enable statisticians to view aggregated results as predictable amidst the local variability of the data. Nevertheless, the design of *ChanceMaker* appeared to support students in articulating situated heuristics such as "the more times you throw the dice, the more even is the pie chart," and to recognise that such a heuristic could be applied to explain the behaviour of various gadgets.

Whereas Pratt's approach led to students initially articulating local meanings before the emergence of global meanings alongside the local, Johnston-Wilder (2006) has shown how the perceptions of students of various ages (the youngest were three years older than those in Pratt's study) shifted back and forth between the local and global, as they attempted to identify whether various types of dice were fair or not. In one-to-one interviews, students rolled each die and recorded the outcome, pausing frequently to reflect on the observed sequence. As the outcomes unfolded, students' attention shifted between a local perspective, in which successive outcomes were seen as disordered and unpredictable, and a global perspective, in which they looked for an empirical distribution of the outcomes. This shift of attention was sometimes rapid and often subtle. In the light of Pratt's study, we noted the direction of the shift with particular interest.

In Johnston-Wilder's study, students were trying to make judgements about the fairness of various different kinds of dice by looking at the observed outcomes. A student's attention was initially focussed in the local perspective, typically looking at short sequences of outcomes, and often seeking patterns in these. Although such patterns might appear by chance, they were typically not sustained as more outcomes were generated. Students often discerned sequential patterns, which they thought might be extended in later outcomes, using them to make predictions about future outcomes. However, these illusory patterns appeared to be a significant distraction for the student trying to make a judgement about whether the dice might be considered to be fair.

Looking at the outcomes through the global perspective involves looking across the space of possible outcomes to consider the frequencies of the various outcomes. The focus of attention is not on successive outcomes, but on the distribution of the outcomes across the outcome space. To see the process through the global perspective involves a different way of looking at the outcomes, and through this perspective it is possible to see an emergent order and pattern in the distribution.

However, the distribution of outcomes can be thought of in two different ways. Firstly, the distribution can be a theoretical probability model which expresses what one expects from a process. Secondly, the observed distribution of outcomes might be viewed empirically as a representation of the underlying probability model. In the case of a student experimenting with, say, a spherical die (a hollow sphere with a hidden weight moving around inside the shaped interior such that it comes to rest in one of six different orientations), the student might expect the die to be fair, beause the spherical appearance gives the die an apparent symmetry. The student might therefore hold in mind a theoretical distribution according to which they expect each outcome to occur equally often. Working with this theoretical prior model of a distribution, such a student might try consciously to control the process of throwing the die to produce an outcome that had not yet been observed in the outcomes so far. In doing so, the student has shifted from the global perspective, in which they were comparing the observed frequency distribution of outcomes with their mental model of what they expected the distribution to look like, to a local perspective in which they look for a particular outcome to occur next.

In interviews with students as they experimented with the different dice, Johnston-Wilder observed that the students' attention shifted between, on the one hand, the unpredictability of the next outcome and the lack of order and pattern in short sequences of outcomes, and on the other hand, the order and pattern that was seen to emerge in an empirical distribution. When the student tried to infer the distribution from small samples, the apparently conflicting information arising from successive small samples appeared to lead the student to make particularly rapid shifts of attention between global and local perspectives.

There are two interesting contrasts between the use of global and local by Pratt and by Johnston-Wilder. Whereas Pratt's work focussed on the emergence of the global out of the local, Johnston-Wilder's observation was one of constant shifting of attention between the local and the global. Secondly, in Pratt's work, the students had available to them the workings box, which came to be seen as a representation (of distribution from our perspective) that could be used to predict behaviour and results, whereas Johnston-Wilder's students could see the dice but did not have access to the associated probabilities. Pratt's students could therefore draw on information about the workings box as well as the generated data, whereas Johnston-Wilder's students had available the data and whatever prior distribution they held for the dice, in the sense of expectations about how it might behave.

This difference could be important because inference is more closely related to the activity of Johnston-Wilder's students. When statisticians make inferences, they attempt to make descriptions of the population (the distribution, or at least statistical parameters such as the mean, which describe elements of that distribution) based on data that have been sampled. Johnston-Wilder's students, in trying to understand what was happening when the die was thrown, were trying to infer something about the underlying infinite probability distribution (at least from our perspective), in some respects the inverse of what Pratt's students were doing. By attending to the workings box, a representation of the modelling distribution, and observing consequential changes in the data, Pratt's students might be expected to see, in Prodromou's (2007) terminology, intention in how the human or computer agent generates the data through the modelling distribution. In contrast, again using Prodromou's terminology, one might expect Johnston-Wilder's students to see the data as targeting the modelling distribution in the way that the model emerges out of the data. However, it is unclear whether the shifting of attention observed by Johnston-Wilder could simply be accounted for by differences in theoretical perspective from that of Pratt, or by methodological issues related to the task differences, whereby target connections are more prone to such shifting than intention connections.

In this paper, we seek to elaborate on the shifting between the local and the global in such a way that we begin to recognise the difficulties that students may have in making informal inferences that connect data to probability distribution. Our approach will be to examine fresh data, arising from a small-scale study in which students were challenged to infer the nature of a single hidden *ChanceMaker* die, given data being generated by the students' manipulation of that die. We aim, through this new elaboration, to guide the design of new resources that would have the potential to support young students' inference-making. In this respect, this paper could be thought of as an extended reflection, marshalling thoughts geared towards the next phase of a long-term design experiment. We have found it useful to extend our corpus of data from the work of Pratt and Johnston-Wilder with some fresh data, collected in an attempt to tease out the subtle differences alluded to above. In the next section, we explain further the basis of the additional data which will inform our understanding of the relationship between the local and the global in statistical inference.

In our discussion of the new data in this paper, we have drawn upon a framework for the Structure of Attention (see, for example, Mason & Johnston-Wilder, 2004). This describes how a person's attention can shift rapidly between different foci, related to different ways of attending. Mason and Johnston-Wilder refer to the following five ways of giving attention to a situation:

- attention on the whole, the global;
- attention on distinctions, distinguishing and discerning aspects, detailed features and attributes;
- attention on relationships between parts or between part and whole, among aspects, features and attributes discerned;
- attention on relationships as properties that objects like the one being considered can have, leading to generalisation;
- attention on properties as abstracted from, formalised and stated independently of any particular objects, forming axioms from which deductions can be made.

(Mason & Johnston-Wilder, 2004, p. 60)

There is considerable evidence that people working to make sense of random phenomena are sometimes attending to what will happen next (immediately), to what has just happened, and to what has been happening over a longer period. They may easily circle around these very rapidly, or focus for a time on one or the other. It is a reasonable conjecture that some people at least are *seeking a relationship* between two or three of these, not always with success. But the overall goal of the instruction is that they see these relationships as examples of properties that can hold in other situations.

Children in the data discussed below run into conflict with the essential unpredictability in the short-term, and the fact that only long-term statistics (summaries of data) are likely to show some relative invariance. Sometimes the children are *discerning details* in the graphs for example (past history) and *seeking relationships* between the past history and the number of occurrences of something. But, in the back of their minds there is a nagging doubt because these 'relationships' are at best approximate.

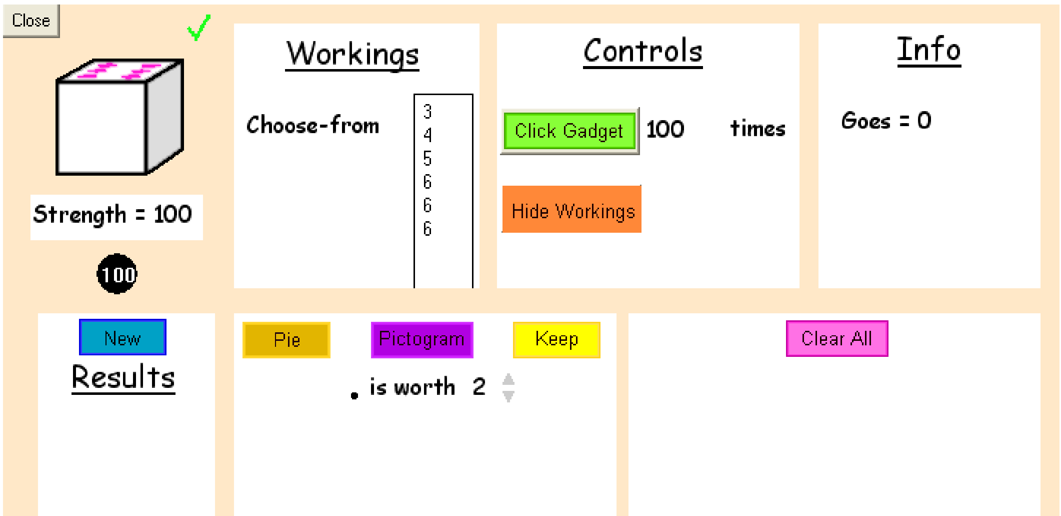
One aim of the tasks is to get learners to restructure their attention from the local to the global, from the details of specific events to the summary statistics of a large number of events. This could be described as *perceiving a property* which is instantiated in the data collected. However, this has an extra wrinkle because summary statistics are never exact, only approximated by instantiations.

In the discussion of the data that follows, we use this way of thinking to *account for* the way that the children often show a somewhat tenuous desire for more trials. Sometimes it is the interviewer who suggests the need for more trials, and sometimes the idea comes from the children. We suggest that, in this study, children seem to be seeking relationships between three distinct features of the software: the 'workings box' (whether they have access to it or imagine it); the role of a 'large number of trials'; and the graphical representation of the outcomes (whether it be 'pie-chart or pictogram').

3. A FURTHER APPROACH TO STUDYING THE LOCAL AND THE GLOBAL IN INFORMAL INFERENCE

Building on the work of Prodromou and of Johnston-Wilder, we designed a small-scale study, which aimed to explore children's thinking-in-change (Noss & Hoyles, 1996). One aspect of the study was to examine in detail students' attention to the local and global when they worked with *ChanceMaker* as in Pratt's original study. However, in order to connect this work to the second aspect of the study, and because of our limited resources, we restricted the children to work only with the die gadget.

For the second aspect, we wished to explore a situation which seemed to bear the hallmark of Johnston-Wilder's work, in that the students were not aware of the underlying probability distribution, and yet had the advantages of simulation, such as the ease of generating large amounts of data as available in the *ChanceMaker* study. In fact, we recognize that in pedagogic tasks which aim to engage pupils in inference, there is a potential gap between the focus of the teacher/designer on reasoning about the population (looking at the population through the sample as in Game 2 in Section 1) and the focus of pupils who may be looking at and reasoning about the sample, without understanding the true nature of the game (as in Game 1 in Section 1). We therefore wanted to design a context in which it would be clear to the pupils that they needed to attend to the sample *in order to* reason about the population. We therefore implemented a simple modification to the original software, which allowed the 'workings box' for the die to be hidden (see Figure 2). Using this modified software (which, for the sake of clarity, we call *InferenceMaker*), we designed a task based on the creation of a 'funny' die which children could explore in order to guess what numbers were on its sides. The software allowed the user to enter any numbers into the workings box, and also any number of numbers, so it was possible to 'make' very unusual dice.



Strength = 100

Choose-from 3 4 5 6 6 6

Click Gadget 100 times

Goes = 0

• is worth 2

Figure 2. In InferenceMaker, it was possible to edit the workings box and then hide it by clicking the "Hide Workings" button. Children were then challenged to infer the configuration of the die by generating data and charts.

We conducted clinical interviews with small groups of 10- to 11-year-old children, in the final year of primary school. We worked with children from a single class group, covering a range of attainment. While the children were working, Camtasia^{TM} software was used to capture all activity on the computer and audio recordings were made of all discussion. Field notes were kept by the researchers.

When working with *InferenceMaker*, after a brief introduction to the software, a funny die was created in secret by editing the workings box. This was then hidden, and the children were challenged to guess what the new die looked like, using any facilities of the software that they chose. To focus their explorations, they were told how many sides the die had (i.e., how many items were in the workings box). Once they had decided on a description, a discussion was held about how the description was arrived at, before the workings box was revealed.

In the ensuing elaboration, we shall first discuss students' attention on the local and global when the workings box was available, as in the original *ChanceMaker* study, and secondly two main themes relating to how the students' attention to the local impacted upon their inferential reasoning when the workings box was hidden. Our focus on attention throughout Sections 4 and 5 is informed by the Structure of Attention framework.

4. THE LOCAL AND THE GLOBAL WHEN CONNECTING THE WORKINGS BOX TO THE DATA

We report first on the activity of Jim and Ivan, as they worked on mending the die gadget (as in the *ChanceMaker* study) for 30 minutes. The boys were presented with a 'broken' die gadget, whose workings box contained too many sixes (see Figure 1). The workings box was visible, but at first the boys paid little attention to it. Most of the initial interaction with this gadget was done by Jim. He approached the investigation systematically, ensuring that each batch of data had exactly the same number of trials. During this work, Ivan was very quiet and said almost nothing. After several experiments with 32 trials in each, Jim reported that "the dice might be a bit wonky if it keeps having six as the most, like the biggest." However, although he had stated clearly that the die was 'wrong', he later commented that "it could just be total luck that it happens to do that.".We continued to press the two boys to suggest ways that they could be more certain about whether there was something wrong with the die gadget, but the boys' only suggestion was to experiment with the 'strength' control, to see whether this made a difference. At this point, Ivan began to contribute more, and after a further set of 32 trials, he quickly (and correctly) stated, "It doesn't make a difference." Jim agreed that this suggested that "there might be something wrong with the dice."

The persistence of the relatively high frequency of sixes over several sets of trials attracted Jim and Ivan's attention. They attended to the detail of the proportion of sixes in each set of trials, and related the frequency of sixes to the frequencies of the other outcomes. The details became, for the boys, a property, which Jim has expressed as "something wrong with the dice." The boys compared what they noticed with their unarticulated expectation that the distribution of outcomes should be uniform.

The boys then turned their attention to the workings box, and Ivan immediately identified that there were "loads of sixes in there," confirming their conjecture. He edited the contents of the workings box, deleting two of the three sixes, to leave one of each digit from 1 to 6. When the boys went on to test the mended die, Jim again took control of the experiments, but the uneven outcomes from the 32 trials left him unconvinced that the die was truly mended.

For clarity in all protocols we have adopted a convention in which we use the word, such as 'six', to represent the cardinal number on the face of the die or in the workings box, and the digit, for example ‘2’, to represent the frequency with which an outcome appears in the workings box or in the results.

> Res: So what do you think now? Do you think the dice is mended or do you think it’s still wonky?
>
> Jim: Well, dices aren’t supposed to do equal amounts of each number, really. Cos they’re dices, but… it’s… it doesn’t give out that much, but… it’s not like having six as the biggest thing any more… We could probably fix it properly if we changed the numbers round on the workings a bit more. And we could make it equal I suppose.
>
> Res: How would you do that?
>
> Jim: Umm… If there were less ones, you could put 2 ones in. Or something. But that would increase its chances by two and then that would make it go wonky, so…

In contrast, Ivan saw no difficulty with the conclusion that the die was now mended. He was certain that the die was working properly, “Because the workings there, everything’s good.”

> Res: Oh, right! But it’s not coming… the numbers aren’t all coming out the same amount. Does that matter?
>
> Ivan: No. Because in real life each one will sometimes be higher.

Even after they had altered the number of sixes in the workings box, the proportion of sixes observed in the subsequent set of 32 trials was still subject to the vagaries of random variation; 32 trials is still a relatively small sample upon which to base a judgement. There is no evidence here that the boys have yet perceived the wider and more powerful property relating to large samples: the Law of Large Numbers. Jim in particular appeared to expect more stability than he found from his relatively small samples of 32 trials.

5. THE LOCAL AND THE GLOBAL WHEN INFERRING FROM THE DATA

In this section, we consider how students worked with *InferenceMaker*, where the workings box was hidden. The children engaged enthusiastically with the task. They typically began by making single throws of the die and looking at the results list, but, with a little prompting, they moved to using the pie chart or pictogram to make predictions about the die as they extended their sample. Their final samples varied in size considerably, and there was no general recognition that a larger sample might be more effective. We shall say more about this in the next subsection, but first we want to discuss the dominant tendency for students to attend to the local.

A recurring feature of the children’s activity was their focus on the changes which occurred in the appearance of the graph as they grew their samples by adding more throws, and the relative invisibility of more stable features.

Rob and Carl had been given a ‘funny’ die with six sides [3 4 5 6 6 6]. They decided quite quickly that the only numbers on the die were three, four, five and six. They also realised that six was occurring most often, and so conjectured that there were 2 sixes (‘the six is 2’). This left them with four faces on the die to fill. As their sample grew, three and five seemed to appear more often than four (though not as often as six). As they added more throws, and studied the pie charts, they explored different ways to describe the die. This conversation took place when they were examining the pie chart, generated from 130 throws (see Figure 3).

Rob: The six is 2 [sides].
Carl: Six is definitely 2 or 3. Three is 1, I think. No four is definitely 1. I think three is…
Rob: It's either three or five, isn't it [that is 2 sides]?
Carl: Three or five, may be 2.
Res: Why do you think six is 2?
Carl: Because it's getting the most rolls.
Res: Ah, why do you think it's 2 and not 3?
Carl: We don't, I think it's either 2 or 3.

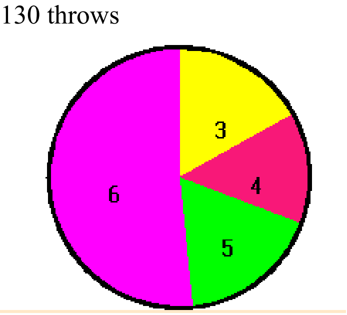


Figure 3. Rob and Carl make informal inferences after 130 throws.

At first sight, it seemed puzzling that the boys did not 'see' that six was taking up a much larger slice of the pie than either three or five, and so could not have been representing the same number of sides of the die. However, their extended discussion showed how their attention was, at this point, focussed on the way each slice was changing as the sample grew. The five-slice seemed to be getting bigger, but was it bigger than the three-slice? Their attention was focussed upon the way that each slice was changing and on the need to fill the appropriate number of 'faces' on the die: They recognised that if there were 2 sixes, and only 4 numbers in total, there must be another duplicate number somewhere on the die.

Throughout their work on the task, Rob and Carl seemed willing to grow their sample with larger numbers of throws, but whether in the hope of seeing stability in the results or simply because this was the only action available to them other than to make a conjecture and settle upon it, we cannot tell. The local and sense-triggered phenomenon of change seems to be at the forefront of their attention, whereas invariance (stability) is hard to detect as a relationship, particularly so because each new graph replaced the previous one as more throws were added. They gave no evidence (as we might have wished) of *perceiving a property* (global) of 'settling down in the long run'. In the end they had a sample of 280 throws before agreeing on their final (correct) description of the die.

We then challenged their confidence in this description by making a new sample of 10 throws. Their confidence proved to be fragile (Figure 4).

Carl: Six is the most
Res: What do you think?
Rob: 2, 2, 2 (pointing to the three sections)
Carl: No
Res: It's the same dice
Carl: 2 sixes, 2 threes and the rest are 1
Rob: Yeah but remember there's a four as well 2 sixes, 2 threes
Carl: 2 sixes, 2 threes and then 1 is five and 1 is...
Res: But you were saying to me that there were 3 sixes and 1 three, 1 four and 1 five
Carl: This is confusing!

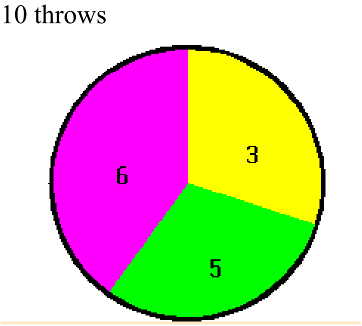


Figure 4. Carl and Rob respond to our challenge based on 10 throws.

Adding ten more throws to the sample produced yet another image, with a small slice for four and a reduced slice for five. When asked which image they believed. Carl favoured the one for twenty throws: "Because the 280 was just getting too stupid, I think, and had too much in." Rob set about making a sample of 140 throws, half way to 280. Carl's dismissal of the larger sample as "too stupid" seemed at odds with his earlier willingness to create more data, but may have reflected his frustration that more data seemed to produce more change rather than more stability. Carl's dismissal of the larger sample might be explained in terms of the way senses work; the senses may have been activated by change and the learner's attention may have been drawn to change. Although Rob and Carl were expecting invariance in the midst of change, as they may have learned to expect from their past experience of mathematics and from learning to make sense of their experiences in daily life, their attention was drawn to change rather than to invariance. They lost sight of the significance of the global perspective (the results of the larger sample of 280 outcomes) as the change arising from what had happened most recently in the smaller sample of 10 outcomes (a local perspective) had attracted their attention.

It was unclear what Rob and Carl were expecting or hoping to see as they added more and more throws to their sample. We conjecture that they were expecting to reach a point where the new, larger sample produced a graph which looked the same as the previous one; that is, where there was no change. However, as even small changes were grabbing their attention, because this is what our senses are most attuned to, they became frustrated. Even though they had drawn an initial inference from a sample of 280 outcomes, their attention was drawn by the change apparent in the subsequent sample of 10 outcomes. They did not have a secure perception of the (global) property that large samples were more informative than small samples, and perhaps assumed that the strategy of taking larger samples was the wrong one.

We now return to the work of Jim and Ivan, whose first encounters with the workings box in the software were described and discussed in Section 4. In what follows, we had used *InferenceMaker* to set up a 12-sided die whose workings box was hidden and contained the following: [1 2 2 2 3 3 3 4 5 6 6 7]. Jim and Ivan were told only that the die had 12 faces. Jim initially took the lead in controlling the software and generating data until they had a pictogram of 100 outcomes (Figure 5). Ivan commented from this graph that there were "loads of twos and threes," but Jim found difficulty in interpreting what this graph implied about the die.

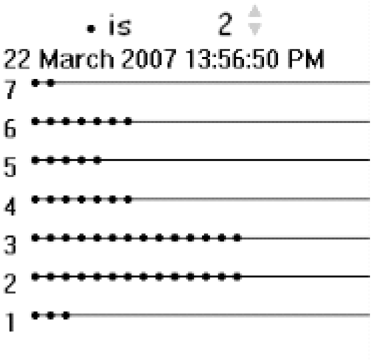


Figure 5. Jim and Ivan generated a pictogram based on 100 throws.

Ivan: Well last time it was all different (*indicating the outcomes*), and it was the same numbers (*indicating the workings box?)*

Res: So what might it have?

Jim: Umm. It might have 2 sevens and 3 ones and … (pause) and 7 sixes…

Jim seemed to have switched to describing the numbers of spots in the pictogram as though they were exactly the numbers of each element in the workings box. This appears to be an unintended metonymy, in which Jim's attention has switched from one detail to another. In contrast Ivan appeared to have a clear idea of how to interpret the pictogram.

Res: Which number do you think that there might be most of on [the] dice?
Ivan: Well, threes and twos.
Res: threes and twos. …So do you think there might be the same number of threes and twos?
Ivan: No.
Jim: Yeah. Well it's going to be more than the fives, fours, sixes, sevens and ones.
Res: OK.
Ivan: But it doesn't have to be the same number of threes and twos because it could be just accident.
Res: So what could we do to be a bit more certain?
Ivan: Do it again.
Res: Do another 70 throws? Or another 100 throws?
Ivan: One hundred.

Ivan stated clearly here that his strategy for understanding what was in the workings box was to collect more data. He clicked the gadget to collect a further 100 outcomes, which were added to the previous 100, and he produced a new pictogram (with a different scale) to display all 200 outcomes (Figure 6). When Jim commented on this graph, he expressed a plan to save this graph, collect a new sample and compare the resulting graph with this one.

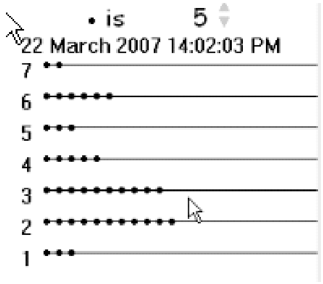


Figure 6. Ivan generated a pictogram for 200 throws.

Jim: If we kept that, and then made another, made a new one and then did another 4 hundred, we could see if there were actually the same problem. I think that there's more twos than threes cos it says (inaudible)… And we've made it more fours… threes… and that's where we can see which is the biggest.

Jim's proposal for taking a new sample, rather than continuing to grow the existing one, may suggest that, like Rob and Carl, he has recognised that growing the sample is not producing the stability he is looking for. His plan to compare samples may be an expression of a shift in his thinking to a more global perspective. Jim produced the new graph of fifty outcomes and displayed it beside the previous graph of 200 outcomes (Figure 7).

However, as the boys contemplated these images, Jim's attention was again drawn to local changes. He noted that the new graph showed more threes than twos, in contrast to the previous one. The Researcher's questions in response to this remark seemed to prompt

Ivan to look at the data in a new way, as he suddenly began to express a complete description of what was in the workings box.

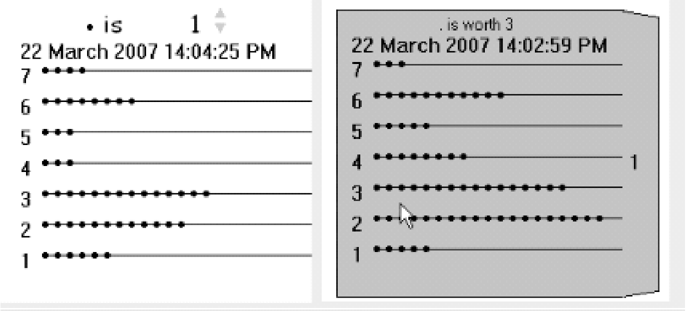


Figure 7. Jim compared a graph of 50 (on the left) outcomes with one of 200 (on the right).

Ivan: Now there's more threes.
Jim: …There's more threes than twos. So that means that that could have been randomisation.
Res: So, do you still think there would be more twos than threes on this dice? Than the other numbers?
Ivan: Yes.
Res: What about the other numbers then? Do you think they would all be the same? Do you think that there'd be the same number of one, and four, and five, and six and seven?
Ivan: I think that there's 2 sixes.
Res: 2 sixes?
Jim: Well five and four could be the same, but, we did think that two and three would be the same but then two got bigger and then three got bigger, so…
Res: We've got 12 places to fill up. So there was…?
Jim: If I divided it by 12…
Ivan: Yeah, I think… I think I am right…
Res: Go on then. What do you think?
Ivan: Cos I think that there's in the workings there's 3 twos, 3 threes and 2 sixes. And that makes 12.
Res: Ahh. Right. …Say it again, 3…
Ivan: 3 threes, 3 twos and 2 sixes. And that all these numbers for one.

This conjecture was recorded on paper for the boys to consider. Ivan's guess here was in fact correct, but he was not certain of it, and he tried to compare recent graphs with those that they had generated earlier. There was a particular difficulty here as the graphs did not record the number of trials that they showed, and early graphs had been derived from collections of outcomes with differing numbers of trials.

Jim decided to collect 150 more outcomes to add to the 50 shown on the latest graph, and when he did so, the new sample of 200 showed a greater proportion of sixes shown than before.

Ivan noted this and wanted to revise slightly his earlier guess. Faced with the dilemma of which guess was more likely to be correct, both boys suggested collecting some more data. Again they collected a sample of 200 outcomes and again they did not consider that the graph was conclusive. Eventually, the interviewer suggested that they might collect a larger sample of say 500 outcomes. Ivan quickly adopted this suggestion and the resulting graph convinced him that his first guess was correct.

There seems to be evidence, in Ivan and Jim's response to the graph in Figure 7, of a shift towards *perceiving the property* that a large sample size gives useful summary statistics. However, when they looked at the next sample of 200 outcomes (Figure 8), their attention was again drawn to change rather than invariance. Perhaps they were seeking too rigid or robust an invariance, in which case there were aspects of the property still for them to appreciate. An important consideration relating to the property of invariance in distribution is what degree of variability can be accepted while still recognising invariance.

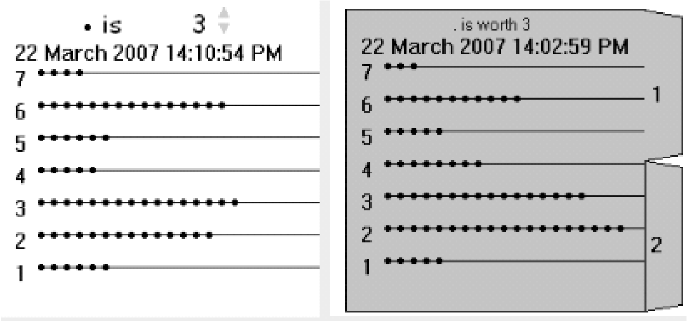


Figure 8. Jim's new sample of 200 throws showed more sixes.

Within our design research approach we adjusted the exact nature of the tasks as we worked with different groups of children in order to explore ways to support their thinking. We tried an approach in which we set a series of challenges for a group of children with dice which were progressively more complex (four sides: [1, 2, 2, 3], six sides: [1, 2, 2, 3, 3, 6], nine sides: [1 2 2 3 3 4 4 4 6]).

Alice, Freya and Bella quickly realised that the four-sided die had only three numbers. After only 12 throws they made a pie chart and recognised that two was occurring more frequently than one or three. Freya expressed this in a proportional way, though it is not clear if she was talking here about the pie chart image, the imagined die, or the workings box.

> Freya: The two is double… there's 2 in it … 2 twos

It was relatively rare for children working on this task to use proportional language to describe the images. Generally they talked in terms of relative sizes, but did not attempt to quantify these differences. This might be seen as evidence of their local, rather than global, focus. When we challenged the girls' prediction by showing them a different sample of 10 throws, the image confirmed their decision.

However, Freya's use of language may point towards a power that is released when representations admit the same language for different ways of perceiving; there can also be unintended metonymies, where attention switches between details (such as attributes), because of similarity of language in both perceptions (in this case, the same number). We asked ourselves what Freya was referring to when she said "there's 2 in it." She might be referring to the pie chart image, or the workings box, or the die. Initially, Freya seems to describe the pie chart image ("the two is double"), but, perhaps because there is a potential ambiguity to this reference, she seems to shift towards possibly referring to the workings box or the die. The use of ambiguous language such as this may offer the potential for the speaker (and perhaps the listener) to slide across from referring to the data representation to describing the die, or even the workings box, and in this way to see beyond the data to the sense in which the data provide insight about the population. Such a switch of attention, prompted by the ambiguity in the language used in trying to express what has been observed, might provide a stimulus to 'see beyond the data' and to begin to infer the nature of the population.

The six-sided die proved a little more challenging, but the girls quickly reached a point where they had a number of different conjectures about the distribution of numbers. As they worked, they seemed at times to accept implicitly that doing more throws was a way to test out their conjectures. The following comes from a point where they have done 33 throws.

Alice: one, six, two, three, three, two
Bella: yeah … that's it
Freya: Are you sure?
Bella: Let's do it again, and if the threes keep coming up…
Freya: It's three … it has to be three

The girls continued to add throws without discussing this strategy, and to make further conjectures about the die. They made a new pie chart after each group of 10 throws, and the local changes in this became the focus of their attention. At the stage shown in Figure 9 the three-slice on the pie was larger than any of the others, and the difference between the two-slice and the six-slice was less pronounced. They were struggling to reconcile the image in front of them with what they knew to be possible combinations in the workings box. At one point, they suggested [1 2 2 3 3 3 6], but realised that this would give 7 sides.

Bella: But if it was [1 2 2 3 3 6] I think that the threes and twos would be the same.
Res: Ah!
Freya: Oh yes
…
Bella: The three and the two would be the same size
…
Alice: It's definitely three and two have got the most numbers and one and six have got the least.

63 throws

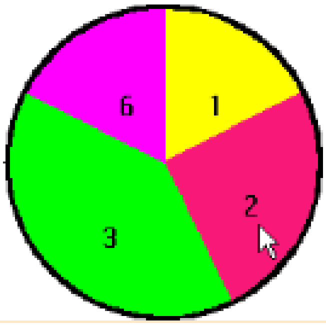


Figure 9. Alice and Bella draw informal inferences from 63 throws.

Once they had decided confidently that [1 2 2 3 3 6] was their prediction, we challenged this by keeping the pie chart shown above, and making a new sample of 10 throws (Figure 10).

Initially this seemed to shake the girls' confidence in their decision, but Bella soon introduced a different perspective, suggesting that she was considering all the graphs they had looked at, not just those immediately in front of them.

All: It's twos!
Alice: It must be I reckon
Freya: Well it's two and three, two and three
Alice: But look at the six there … it's six and two

Bella: But there's …most of the sixes have been quite low, but most of the threes have been big … if you look at the two it's been right in the middle and it's…

Freya: On that one … on that one it's six and two

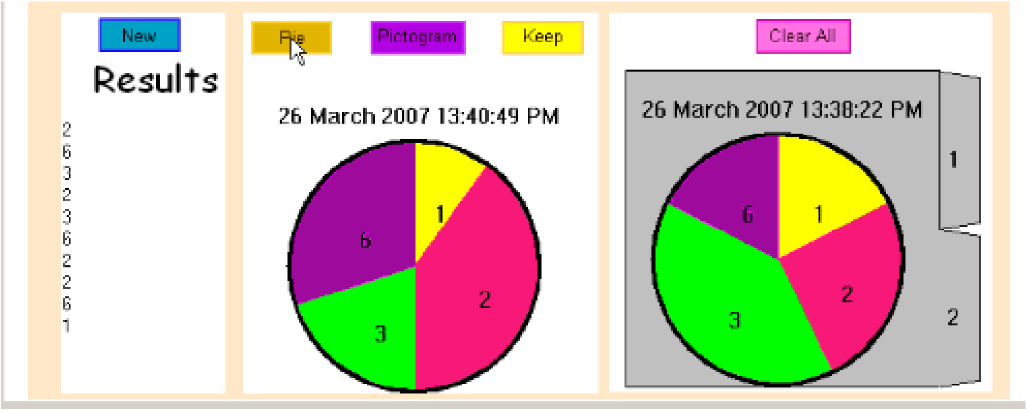


Figure 10. We challenged Bella and Alice with a pie chart based on 10 throws (on the left).

As in the example from Jim discussed earlier, this might suggest that Bella is able at this point to take a more global perspective and see similarities across the set of graphs, rather than focussing on changes between them. Despite Bella's insight, however, the girls continued to be influenced by the new image in front of them, and to make new conjectures about the die. However when asked directly about whether the different number of throws that produced the two graphs might make a difference, the girls agreed that it would, and decided to add more throws to the new sample. They quickly obtained an image which reinforced their confidence in the prediction of [1 2 2 3 3 6].

After their success with the six-sided die, the girls were keen to tackle the next challenge, but found the nine-sided die [1 2 2 3 3 4 4 4 6] much more difficult. Throughout their work they implicitly increased the sample size to get a clearer picture, but actually never went above 100 throws (which they considered to be a very large number). A sample this size had been large enough to show some stability for a six-sided die, but was not sufficiently large to fully explore the nine-sided die. Their strategy was to save graphs, and then begin new samples, but this became complex as they lost track of the sample size for each graph. They continued to struggle between descriptions of the distribution, which seemed to match the images in front of them (but might contain more than 9 numbers), and those which they knew were possible for a nine-sided die, but which did not fit comfortably with the graphs.

6. DISCUSSION

6.1. SAMPLING STRATEGIES IN INFORMAL INFERENCE

When children were working on the task of guessing what the 'funny' die looked like, we saw them use two different strategies which were supported by the software. Having made a graph from a particular number of throws which was inconclusive, they either 'grew' their sample by adding more throws, or they extended the data available to them by beginning a new sample. The latter strategy was supported by a facility in the software to keep a number of graphs, and return to them. In one sense both of these strategies involve increasing the sample size, but the two experiences they provide are very different.

In passing, it is worth mentioning that what is involved in ‘growing the sample’ here is rather different from the activity described by Ben-Zvi and Sharett-Amir (2005), in which children widened the scope of their data collection from a small group of friends, to a whole class, to several classes, and so on. In their activity, it may not be entirely clear, from the children’s perspective, whether it is the sample or the population which is ‘growing’, but the data collected in each iteration can clearly be distinguished. In the case of our task, growing the sample clearly involves collecting ‘more of the same’ data, but the nature of the sample remains the same.

An important question for our thinking about the future development of the task design is how the two different experiences of adding to the sample and taking a new sample impact on children’s local and global thinking. It is clear from our data that neither experience leads easily to the recognition that larger samples are more reliable in providing an image of the distribution; that is, to an appreciation of the Law of Large Numbers. In order to understand this somewhat surprising outcome we need to conjecture about what the children might be *expecting* to see.

We conjectured earlier that Jim was expecting (or hoping) to see invariance in terms of consecutive graphs which stayed much the same as he added to the sample. Given the natural tendency to focus on even small changes, this expectation is almost certain to be confounded, even within sample sizes much greater than those the children were prepared to explore. Seeing repeated change seemed to make some children distrust taking larger samples. Of course taking repeated samples also produced images which reflected change, but possibly the experience of looking at several similar graphs allowed Jim and Bella, if only fleetingly, to gain some sense of the overall stability of the patterns.

However, there is perhaps another way to think about what the children were expecting to see: that is, to see a ‘clear’ pattern. An image in which the sizes of the pie slices, or the lengths of the pictogram bars, were clearly ‘in proportion’ (i.e., some two or three times the size of others) might have proved very convincing, regardless of the size of the sample. Indeed, one group of children did spend time adjusting the scale of the pictogram in order to try to produce such images, with the smallest portions represented by one bead. When children in our study were adding to the sample, they may have been expecting the graph to ‘settle down’ to a clear image, and were frustrated when this did not appear to happen within the sample size they used.

6.2. INFORMAL INFERENCE AS EMERGENCE TOWARDS A TARGET

We are struck by how rarely during the *Guess-my-dice* game in *InferenceMaker* the children referred to luck or chance. Jim was one exception when he was trying to explain to himself the number of observed sixes without abandoning the idea that the die was ‘fair’. It has been well documented how people often do not make sense of phenomena through a stochastic model (as in Konold’s, 1989, outcome approach) or avoid facing the nuances of probability by regarding everything as equally likely, just a matter of chance (as in Lecoutre’s, 1992, equiprobability bias). However, we believe that neither of these interpretations quite fits how the children were trying to infer the nature of the die. We think that these students were trying to see through the data in order to identify the die. Their approach was consistent with Prodromou’s (2007) target connection from the data-centric to the modelling distribution, in which emergence rather than probability is the relevant model.

This is somewhat in contrast to the results observed in Johnston-Wilder’s study, where students who were a little older (aged 13 to 18 years) experimented with physical dice; these students therefore did not have such easy access either to larger samples or to graphical summaries of the aggregated data. In Johnston-Wilder’s study, the emergence of the modelling distribution was not such a salient feature for the students, and their attention was not so readily drawn to it. Instead, the students were most concerned with judging whether or not the dice that they were using was fair.

6.3. INFORMAL INFERENCE AS THE SEARCH FOR INVARIANCE AMIDST LOCAL CHANGE

Emergent phenomena involve the actions (and often interactions) of many agents at the local level, resulting in the formulation of identifiable patterns at the global level. In order to discern a modelling distribution in *InferenceMaker*, students first need to attend to aggregated data rather than to individual outcomes. They need to focus upon frequencies, and eventually upon relative frequencies, and to pay attention to the pattern of distribution of these across the outcome space, rather than looking only at changes in a single relative frequency from one sample to another. Appropriate graphing tools might support the student in attending to the distribution of relative frequencies. Once a pattern is discerned, and a possible configuration for the die has been conjectured, then this conjecture needs to be tried to see what patterns of outcomes it will produce. The coordination of attention to each of these agents in order to discern the invariance of an underlying emergent distribution, when each manifestation of the distribution in a sample is different, requires several steps, each of which might be supported by developments to the software.

In trying to make a connection from data-centric to modelling distribution, the students needed to identify the trend, a global pattern, that might be emerging from the data. However, our study shows clearly how students’ focus of attention, when using *InferenceMaker*, tends to be on the local. Rob and Carl focussed on how the five-slice seemed to be getting bigger, rather than on the dominant size of the six-slice. Jim tended to focus on how there had been more two’s but then there were more three’s. Alice, Bella and Freya constantly referred to the changes in the new pie chart compared to the previous pie chart. This attention to the here and now, rather than the aggregated longer-term pattern, characterised the activity throughout our trials. Ivan was a clear exception here, who seemed to have a deeper understanding from the start, although even he had required some prompting to consider a sample size of larger than 200 trials. Bella also showed some evidence that she was thinking about the images presented over several graphs. There is plenty of evidence that the students wanted to find invariance but were constantly frustrated because all they could see was change. In Mason’s terms, the students were unable to hold the same wholes that more experienced statisticians might do because the properties of invariance were constantly hidden by the tendency to perceive change. The task of identifying the invariant properties was made even more difficult because invariance in the aggregated whole is not absolute invariance but relative invariance; to identify the property of relative invariance, one has to notice how the changes in proportion becomes less significant as the sample gets larger.

6.4. INFORMAL INFERENCES ARE OFTEN MADE ON SMALL AMOUNTS OF DATA

We are struck by an apparent paradox that students wanted to generate more and more data, as they were slow to feel confident about their conclusions, and yet they placed no greater confidence on inferences based on large amounts of data than those made from small amounts of data. Konold (1995) has in the past made the point that data are not forceful in persuading people, though our students did seem to choose to generate more data.

We believe this paradox might be resolved from the perspective of emergence. The students were looking for stability at the wrong level. They hoped, in their search for invariance, that by collecting another batch of data, they would get the same pie chart or pictogram as the last one. We have discussed above how different strategies for collecting the additional data shaped the attention.

Inevitably, whichever way they sought to collect data, there would be change which would attract attention away from their search for invariance at the global level towards the local level where invariance could not be found. How that data were displayed could also make a difference. Pictograms tended to emphasise differences between the lengths of bars in the case of dice with fairly uniform distributions but would be effective in showing up patterns in more distinctive distributions (such as would be the case with the frequencies of the totals of two dice). Pie charts tended to be ineffective in showing up such patterns but reduced the apparent differences between sectors.

Nevertheless, even when using pie charts there was a powerful attraction towards local change. And so, in practice, the extra data provided only more complexity. Carl pointed out that more throws were “just getting too stupid” because they “had too much in.” Somehow, Carl needed to focus at the global level to find the stability he wanted.

Eventually, the students would sometimes see no added value in generating yet more data and would be content to make their inference on what we might regard as flimsy evidence.

7. CONCLUSION

When we argue that young students’ naïve informal inference is emergence-related (rather than based on chance), focussed on the local, and made on small amounts of data, we do not present these as misconceptions to be eradicated. Rather, we see these findings as identifying students’ starting points, informing how we should be building new designs for tasks and learning environments which will offer to students experiences that may enable them to construct more sophisticated meanings for informal inference out of these relatively naïve conceptions. And we do not see this as a forlorn hope.

We believe that, although inference involves making a connection from the data-centric distribution to the modelling distribution, this connection is supported by the intention connection in the opposite direction. We conjecture that giving students the experience of mending gadgets before asking them to infer the nature of the die may be a necessary experience to enable students to understand more deeply the connection from data to modelling distribution. Seeing that random mechanisms can generate many different looking pie charts when the data are limited may be another vital experience. For comparison, perhaps the students needed to see the stable patterns generated by a large number of throws. At the outset, we worried that the students would simply throw the die 1000 times and immediately infer the nature of the die. However, this simple strategy was not available to them because they were not paying attention to the global level of emergence. Indeed, it is likely that they had no global resource such as the Law of Large Numbers available to them.

But suppose they did have. Would this mean that the students were in a powerful position to make informal inferences? Possibly not. Then, it would be interesting to explore the level of confidence students place in their inference if they are allowed only a limited number of throws. How would we support changes in thinking towards an abstraction, which might be schematised as "the more data we have, the more confident we can be in our inferences"?

In considering a new design for *InferenceMaker*, we recognise the key objective is to support students in their attempts to observe global relative invariance in the midst of local change. Such an aspiration leads us to consider (i) increased support for systematic recording and reflection, and (ii) functionality for exploring the behaviour of conjectured die configuration for comparison with the behaviour of the unknown die. More specifically

- We recognise the need to enable systematic recording of the students' conjectures about what the die looks like at any particular time through entering the conjectured sides into a blank 'die' with the correct number of sides.
- We would then consider allowing experimentation with the conjectured die for comparison with what the unknown die has generated. In effect, the students would be creating a workings box and generating results from it, thus enabling the possibility of intention connections from the modelling perspective to the data-centric perspective without losing contact with the challenge of finding the configuration of the unknown die.
- We would consider the provision of graphing and summative tools to enable students to compare the conjectured die with the unknown die. (For example, we can imagine deploying graphing and modelling tools of the type available and becoming available in Tinkerplots$^{\text{TM}}$, Konold & Miller, 2001.)
- Finally, we believe there would be value in enabling better recording of the number of throws with easy cross-reference to the display generated.

From the research perspective, it would be valuable to register what appears to be the focus of the students' attention, local data or global effects. We would therefore want to place more attention on recording aspects of body language, gesture, and where students appear to be looking to augment the Camtasia records.

We reconsider the differences described at the start of this paper between the perspectives of local and global thinking presented in the previous studies of Pratt and Johnston-Wilder. The theoretical lens of Structure of Attention may allow us to see these as products of the available technology and task design, rather than as more fundamental discrepancies. Pratt's students, unlike Johnston-Wilder's, were not able to focus on the sequential outcomes of throwing the dice. Although this is available on the screen, the layout makes it difficult to see, and when using the facility to generate data quickly through groups of trials, the individual results appear too quickly to allow attention to focus on each individually. This might be seen as an advantage, because students are unlikely to become distracted by local sequential patterns, but it may also disguise a potential opportunity to move between local and global perspectives. Similarly, Johnston-Wilder's students did not have access to the graphing facilities provided in Pratt's software, which, although they provide tools, may also serve to disguise the growth of the sample, because each graph may appear to be 'the same size'.

As a final reflection, we return to Makar and Rubin's analysis (2007) in which they identified aggregate thinking and sample size as two important components of informal inference. Through the lens of Structure of Attention, we have seen how our students, aged 10-11 years, were drawn to local variation and often the invariant characteristic of relative frequency, apparent in aggregate thinking, was obscure to them. By being aware of the focus of the students' attention, we have not only begun to appreciate why inference may be such a problematic area but also how the design challenge should begin to respond.

REFERENCES

Ben-Zvi, D., & Sharett-Amir, Y. (2005). How do primary school students begin to reason about distributions? In K. Makar (Ed.), *Reasoning about Distribution: A collection of studies. Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-4).* [CDROM, with video segments]. Brisbane, Australia: University of Queensland.

Ben-Zvi, D. (2006). Using Tinkerplots to scaffold informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D1_BENZ.pdf]

Camtasia Studio (Version 6.0) [Computer software]. Okemos, MI: Techsmith Corporation.
[Online: http://www.techsmith.com/camtasia.asp]

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, *32*(1), 9-13.

Johnston-Wilder, P. (2006). *Learners' shifting perceptions of randomness.* Unpublished doctoral dissertation, Open University, UK.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6*, 59-98.

Konold, C. (1995). Confessions of a coin flipper and would-be instructor. *The American Statistician*, *49*(2), 203-209.

Lecoutre, M. P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics, 23*, 589-593.

Makar, K., & Rubin, A. (2007, August). *Beyond the bar graph: Teaching informal statistical inference in primary school.* Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5), University of Warwick, UK.

Mason, J., & Johnston-Wilder, S. (2004). *Fundamental constructs in mathematics education*. London: RoutledgeFalmer.

Noss, R., & Hoyles, C. (1996). *Windows on mathematical meanings: Learning cultures and computers*. London: Kluwer Academic Publishers.

Pfannkuch, M. (2006). Informal inferential reasoning. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/6A2_PFAN.pdf]

Pratt, D. (2000). Making sense of the total of two dice. *Journal for Research in Mathematics Education*, *31*(5), 602-625.

Pratt, D., & Noss, R. (2002). the micro-evolution of mathematical knowledge: The case of randomness. *Journal of the Learning Sciences*, *11*(4), 453-488.

Prodromou, T. (2007). Making connections between the two perspectives on distribution. In D. Pitta-Pantazi & G. Philippou (Eds.), *Proceedings of the Fifth Conference of the*

European Society for Research in Mathematics Education (pp. 801-810). Larnaca, Cyprus: University of Cyprus.

Prodromou, T. (2008). *Connecting thinking about distribution*. Unpublished Doctoral Dissertation, University of Warwick, UK

Prodromou, T., & Pratt, D. (2006). The role of causality in the coordination of two perspectives on distribution within a virtual simulation. *Statistics Education Research Journal*, *5*(2), 69-88.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ5(2)_Prod_Pratt.pdf]

Smith, J. P., diSessa, A. A., & Rochelle, J. (1993). Misconceptions reconceived - A constructivist analysis of knowledge in transition. *Journal of Learning Sciences, 3*(2), 115-163.

Konold, C., & Miller, C. (2001). Tinkerplots (version 0.23) [Data Analysis Software] University of Massachusetts, Amherst (USA).
[Online: http://www.keypress.com/x5715.xml]

DAVE PRATT
Institute of Education
University of London
20 Bedford Way
London
WC1H 0AL

STATISTICAL INFERENCE AT WORK: STATISTICAL PROCESS CONTROL AS AN EXAMPLE

ARTHUR BAKKER
Freudenthal Institute, Utrecht University & Institute of Education, University of London
a.bakker@fi.uu.nl

PHILLIP KENT
Institute of Education, University of London
p.kent@ioe.ac.uk

JAN DERRY
Institute of Education, University of London
j.derry@ioe.ac.uk

RICHARD NOSS
Institute of Education, University of London
r.noss@ioe.ac.uk

CELIA HOYLES
Institute of Education, University of London
c.hoyles@ioe.ac.uk

ABSTRACT

To characterise statistical inference in the workplace this paper compares a prototypical type of statistical inference at work, statistical process control (SPC), with a type of statistical inference that is better known in educational settings, hypothesis testing. Although there are some similarities between the reasoning structure involved in hypothesis testing and SPC that point to key characteristics of statistical inference in general, there are also crucial differences. These come to the fore when we characterise statistical inference within what we call a "space of reasons" – a conglomerate of reasons and implications, evidence and conclusions, causes and effects.

Keywords: *Statistics education research; Context; Evidence; Hypothesis testing; Space of reasons*

1. INTRODUCTION

Statistical inference involves drawing conclusions that go beyond the data and having empirical evidence for those conclusions. These conclusions have a degree of certainty, whether or not quantified, accounting for the variability that is unavoidable when generalising beyond the immediate data to a population or a process. This is in line with Makar and Rubin's (2007) analysis that key ingredients of statistical inference are generalisations (conclusions beyond the sample data), data as evidence, and a probabilistic language. An important rationale for characterising statistical inference in the workplace in this paper is that such a study may indicate which types of statistical reasoning students might later need as employees. Our work-based analysis can inform discussions about what students should learn in statistics education and may complement recent school-based research into informal statistical inference (Bakker, Derry, & Konold, 2006; Ben-Zvi, 2006; Pfannkuch, 2006; Rubin, Hammerman, & Konold, 2006). In particular, this workplace research points to types of statistical inference that are not typically addressed at the secondary school level and yet can be useful to employees.

Statistics textbooks traditionally make a distinction between descriptive and inferential statistics to stress that students should not too easily jump from conclusions about samples to conclusions about a population. However, the distinction also leaves important types of statistical inference unaddressed: types of statistical inference (whether formal or informal) from samples to populations or processes that are different from the well-known and commonly taught inferential techniques of hypothesis testing and confidence interval estimation. Only a few students learn these inferential techniques while research in workplaces shows that many will need to draw conclusions about a process from a sample (e.g., Noss, Bakker, Hoyles, & Kent, 2007; Smith, 1999). For example, when monitoring and improving production processes, employees typically with little formal education are routinely supposed to draw conclusions from samples about the production process for which they are responsible.

Our research in workplaces suggests that few non-graduate employees need to interpret results stemming from hypothesis testing or confidence interval estimation, and even fewer produce such results. These formal techniques are mostly not at employees' disposal, and even if they are, it is often not possible or cost-effective to use them. What they need is to draw conclusions from samples with relatively simple techniques, and generally to base decisions on an awareness of the uncertainty that comes with generalising to a population or process. In other words, learning descriptive statistics (and even exploratory data analysis) does not suffice for the majority of students, while the aforementioned inferential techniques are, as taught in current curricula, beyond the scope of the vast majority.

The goal of this paper is to characterise a type of statistical inference required in many work settings, and we do so by analysing an example of a widely used statistical technique in which statistical inferences are made: statistical process control (SPC). Because the theory of SPC has some similarities with sequential hypothesis testing (more than, say, confidence interval estimation) and because hypothesis testing is a better known type of statistical inference within the educational world, we compare SPC with hypothesis testing. The central question addressed in this paper is therefore *How is the statistical inferential reasoning ideally involved in SPC similar and different from statistical inferential reasoning involved in hypothesis testing?*

To address this question we draw on data collected in our research into the mathematical and statistical knowledge required by intermediate-level employees in various industrial sectors (Tehcno-mathematical Literacies in the Workplace Project, 2003-2007). Such employees are typically non-graduates who may be working in manufacturing as skilled operators or supervisory managers.

After discussing the key ingredients of our question – statistical inference and SPC – we describe the origin of our empirical example and illustrate characteristics of statistical inference as observed in SPC. Last, we discuss the contribution that we think this endeavour has made to the study of statistical inference at work, the limitations of our exploratory approach, as well as potential implications for workplace training and school education and research.

2. THEORETICAL BACKGROUND

2.1. STATISTICAL INFERENCE WITHIN A SPACE OF REASONS

Our experience is that within the statistics community, “statistical inference” mostly connotes formal inferential techniques. However, we use the term inference here in its general sense of drawing conclusions, including the possibly tacit reasoning processes that precede and support the explicit inference from a premise to a conclusion, a prediction, or a conjecture. The term not only includes deduction and induction, but also *ab*duction. Abduction is inference to an explanation, a method of reasoning in which a hypothesis is formed that may explain the data. For example, 8-year-old students in Paparistodemou and Meletiou-Mavrotheris’ (2007) study sometimes came up with abductive conjectures that would explain the data rather than inductive conclusions from the data – contrary to the teacher’s and researchers’ expectations. A similar observation is reported by Zieffler, Garfield, and delMas (2007) for college students.

In search of the delimitations of what counts as statistical inference, we ask the following question: Is the calculation of the means of two samples a statistical inference? In our view, this depends on the reason why they are calculated. For example, they might be calculated to know the difference between the two means in relation to the variation of the two samples. The ratio of this difference to a measure of variation (say, SD) can help us conclude whether the difference is big enough to be likely caused by a difference between two populations from which the samples were drawn. In this case, the very fact of attending to the calculation of the means and difference arises due to specific reasons related to populations.

One way to put it is that the attention to and choice of calculations take place in the “space of reasons” within which people act and think, where “reasons” refer not only to reasons in the strict sense but also to implications, evidence, conclusions, goals, purpose, utility, and our knowledge of causes and effects. For the philosophical background of this notion, originating in the work of Wilfrid Sellars, we refer to McDowell (1996). Our intention behind using this technical terminology is not only to recognize that it is impossible to provide a complete description of any particular context in which statistical inference takes place, but also to recognize that contexts involve not just material but also ideal elements, such as reasons.

To characterise statistical inference at work, it makes sense to make a brief comparison between school and workplace settings, which give rise to different spaces of reasons. At school, contexts are often used to learn about statistics, whereas in the workplace, statistics is more likely to be used to learn about the context. Paraphrasing a famous quote by Steen (2003, p. 55) we can observe that the workplace makes sophisticated use of elementary statistics whereas in the classroom we encounter elementary use of sophisticated statistics. Seminal research into the mathematics used on the street (Nunes, Schliemann, & Carraher, 1993) or in supermarkets (Lave, 1988) led to increasing popularity of situated cognition and socio-cultural theories. In the light of such research, it is likely that context plays a different role in statistical inference learned at school than in statistical inference used in workplaces.

At work, statisticians and practitioners using statistics do not lose the context of an investigation out of sight when using statistical techniques (Wild & Pfannkuch, 1999). More generally, Noss and Hoyles (1996) have introduced the notion of “situated abstraction” to capture both the process and product of situating abstract knowledge in real-life situations such as workplaces. Situated abstractions gain their meaning not only from the mathematics from which they stem but also from the context in which they are used. Ethnographic studies of workplace situations by the same authors, for example of nursing (Noss, Pozzi, & Hoyles, 1999), show how mathematical and contextual meanings such as those of "average" are fully integrated with the particular purposes for which they are used – in this case monitoring the blood pressure of critically ill patients.

These studies show that a dichotomy between statistics and context is problematic (see also Cobb, 1999), especially in workplace situations. One difficulty with the concept of context is that it is not well defined. It is often used to indicate the location or setting in which theoretical ideas are used, but that is a too restrictive connotation for our purpose of characterising statistical inference at work. Given this paper's focus, we will not try to define "context" but rather we will attend to the space of reasons in which statistical inference takes place. In this way, we hope to overcome the dichotomy between statistics and context that seems so deeply engrained into our thinking.

2.2. STATISTICAL PROCESS CONTROL (SPC)

The question addressed in this paper is how statistical inference involved in SPC is similar or different from hypothesis testing. We assume readers are familiar with hypothesis testing but perhaps less so with statistical process control (SPC). We therefore briefly characterize SPC (Caulcutt, 1995; Oakland, 2003) before we describe the origin of the empirical example of SPC (Section 3) and analyse it (Section 4).

SPC is a process improvement technique deployed in many industrial sectors. It is typically used in situations where variability in items produced (or services offered) has to be minimal and key performance indicators need to be very close to a specific target. Measurements of the products can be plotted on an SPC chart so as to monitor the location and variability of the production process.

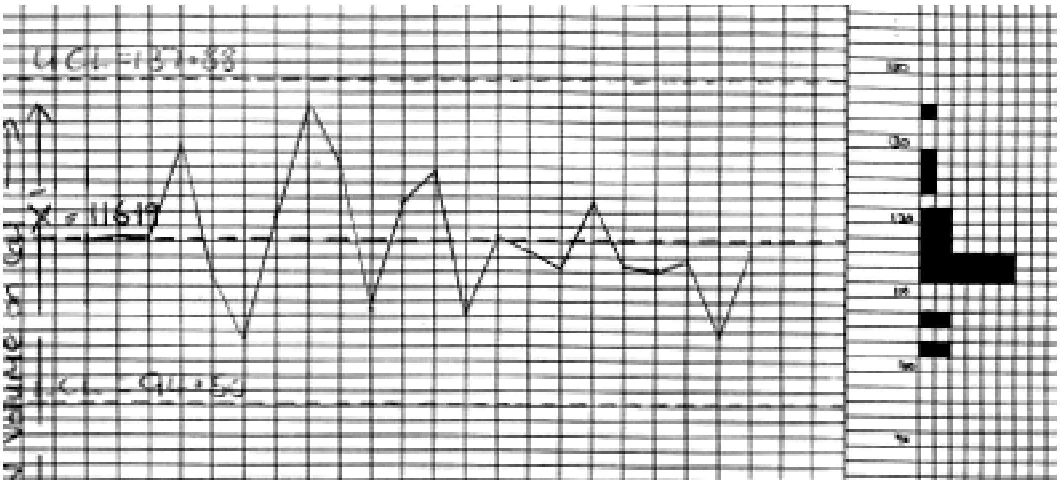


Figure 1. Part of an SPC chart on airtightness of cars

Figure 1 is part of an authentic SPC chart we collected in a car company. It shows the airtightness of a series of cars being tested as they came off the production line. On the left is the control chart with individual measurements; on the right a "sideways histogram" that is used to monitor the distribution of the measurements. The mean of this part of the process is 116.19 (in some unit of pressure that the employee we interviewed could not tell us). The dotted lines below and above the mean line are called the lower control limit (here LCL = 94.55) and the upper control limit (here UCL = 137.88), respectively. These *control limits* are defined as follows: the lower control limit (LCL) is the mean – 3 SDs; the upper control limit (UCL) is the mean + 3 SDs. The use of these control limits is based on two assumptions:

1. The particular measure of items produced is distributed normally (this might be the means of subsamples) – this is checked with the sideways histogram,
2. data are independent and identically distributed.

If the assumptions are fulfilled we can assume that about 99.7% of all measurements will be within +/- 3 SDs of the mean. (Various techniques have been suggested for approximating this standard deviation calculation in a manner accessible to the employee on the production line.) Control limits are also called *action limits*, because non-conforming points (i.e., points outside the control limits, or exhibiting certain trends) are reasons for action. The action usually involves just checking, but sometimes goes on to include adjustments of settings or even stopping the production process for detailed investigation.

The assumption about 99.7% being between the control limits only holds if the mean of the process does not change, that is if there is only *common-cause* (random) variation. However, any *special cause,* such as the supply of different sealing material or a problem with the machine or measurement system, can lead to the process drifting off target. In that case, the cause needs to be identified and addressed. There are probability-based rules for detecting potential trends in such charts. For example, the chance of seven consecutive points on either side of the mean has a probability of only $2\times(1/2)^7 = 1/64 = 0.016$, an occurrence that is therefore unlikely to have been caused by random common-cause effects (this rule can be seen as a binomial hypothesis test). In such cases, one assumes that it signals a special cause (to be identified by abductive inference). In this way, control charts can be useful to detect deviations from the normal or target situation and remove the special cause before the process creeps out of control. Other rules on unnatural patterns include the following: One point outside of control limits; a trend of six points in a row increasing or decreasing; 14 points in a row that alternate up and down.

The idea of using control limits is that they predict process variation so one can stay well within *specification limits*, limits that are imposed by law, by customer requirements, or by senior managers and engineers in the company. Even if data points are occasionally outside the control limits, they will then not surpass any of the specification limits (the spec limits are not shown in Figure 1). If the control limits are well within the spec limits, the process is said to be stable and capable. If the process is stable, control limits can be calculated in preliminary studies, after which the process only needs to be monitored. These are in short the theoretical ideas behind the industrial statistics of SPC or, in other words, *part* of the space of reasons in which SPC takes place.

3. AN EMPIRICAL EXAMPLE OF STATISTICAL INFERENCE IN SPC

3.1. ORIGIN OF THE EXAMPLE

The example that we use to compare statistical inference involved in SPC with hypothesis testing stems from the *Techno-mathematical Literacies in the Workplace* research project. The project goals were to identify the mathematical and statistical knowledge required by intermediate-level employees in financial and manufacturing sectors, and based on these to design learning opportunities that would support employees in developing such knowledge. In manufacturing companies (pharmaceuticals and packaging) we have identified SPC as an important technique that is widely used and yet difficult for employees to understand and use due to the statistical knowledge required

(Hoyles, Bakker, Kent, & Noss, 2007). Contact with a high-level manager from a car company gave us the opportunity to investigate how process improvement techniques were actually used and trained on the shopfloor in automotive manufacturing.

We spent 18 researcher days in this particular car company. First, we interviewed the manager in charge of process improvement as well as the SPC experts and their managers. Shopfloor employees explained to us their control charts on the shopfloor. We also attended and evaluated their SPC course. Moreover, we distributed a questionnaire to twelve course participants (ten responded), and carried out three one-hour follow-up interviews with participants. Some of the operators, shift leaders and course participants had had little or no formal education since they were 16. Our data collection in this company included audio recordings, workplace artefacts such as SPC charts, notes made during training courses, a questionnaire, and trainers' PowerPoint presentations. In collecting data, we made sure in our interviews that we obtained different views of the same workplace activity, from the viewpoints of shopfloor employees, shopfloor supervisors, trainers, more senior managers, and statistical consultants.

In analysing data, we triangulated interpretations of the raw data sources (audio transcripts, photographs of workplaces, artefacts such as graphs) amongst the project team. We have also carried out design-based research so as to enhance existing SPC training, but will not report on it here.

3.2. AIRTIGHTNESS OF CARS

Our empirical example of SPC stems from an area of the production process where the airtightness of cars that are almost ready to leave the factory is checked. Kevin, an operator with no formal education since he was 16, is responsible for this. We cite him to sketch part of the space of reasons in which SPC is used and to give an example of non-statistical inference in terms of causes and effects:

> If the car is too airtight you will get a problem with the windows misting up all the time; also the doors will not shut. You need to lose some sort of air otherwise the door is just so airtight you would have to run at it and give it a good push.

To check whether the airtightness is within specification, Kevin blows air into the car, which is measured in cubic feet per minute, and he reads off the pressure this causes. Mostly the measurements are fine, but occasionally they are out of specification. This is where contextual reasoning is used: "As soon as I turn the gauge on it kicks in normally at around 60-70 Pascal. If it kicks in at around 30 I know full well that we have got a big leak somewhere." (Trying to understand what the numbers meant here was challenging to us, because he kept using different units for both air speed and pressure such as cubic feet per minute, litres per second, Pascal, weight per cm^2.)

Kevin has learned the probability-based rules on trends and patterns, but does not know the probabilistic origin of them. Applying the seven-in-a-row-above-or-below-the-mean rule, he faced particular cases where data points were too high. Abductive reasoning was used to explain the data running high: The sealing material from a particular batch turned out to be different—a special cause—but there was no reason to stop the process. The data points were not outside the control limits and the cause was found. If customers complained, the story would have been different, of course. Having sealing material leading to a slightly different air pressure can be seen as a constraint, and such a constraint can be framed as a reason that is dominant over others (in many situations one would try and bring the process back to the target line again).

Like hypothesis testing, SPC leaves room for two types of errors. The first is that non-conforming data points or trends are observed in the control chart whereas nothing is wrong with the process. Purely by chance this happens regularly: Even if the probability of each individual unnatural pattern is smaller than, say, 0.05, the probability that at least one of about ten such rules finds a pattern is much bigger. The second type of error is that the data points do not show anything special whereas there is something wrong. Kevin gave us one example:

> My main failure at the moment is on estates [station wagons] on the left hand rear wheel arch. (…) In two of the cases the car was not out of spec but I was going round doing my checks to make sure there were no unusual leak paths and we found that they had missed some sealing. So although the car was in spec I raised an AP [Assigned Person, who is responsible for doing the investigation] and still got all the investigations done because there was an unusual leak path.

Thorough knowledge of the work process is required to interpret the implications of certain observations, for example how to measure and what to do about the problem. For example, certain leaks are allowed whereas others are not:

> You just have to go round [the car] and check [the air] is all coming out the normal places, like your door handles. Every door handle has got a massive air leak on, so you allow for that. At the bottom of your windows they let a lot of air out, but then when you start going to your wheel arches and underneath the car, there are certain places where maybe a plug is missing or you have a robot sealer skip.

Kevin knows that the most likely cause for a robot to skip a seal is when colleagues have switched it off and then back again; the robot then always skips a seal. If this is detected then a judgement has to be made whether the car should be sent back for an extra seal; this judgement is most likely made on both contextual and statistical grounds.

From such examples we were convinced that Kevin knew the process very well: He knew what to look for, what might cause it, what implications it has, what to do about the problem and so on. In terms of the space of reasons involved, he was aware of many reasons and conclusions, causes and effects that were linked by—we think—the right inferential relationships. However, when statistical issues were involved he felt less comfortable, for example when interpreting the control limits he and his colleagues received from a central office:

> What [the office] actually said to us is that it should have been 120 [weight per] cm squared +/- 3 sigma. It is very, very complicated because they gave us a lot of specs [actually control limits] to work on and it did not mean a lot to anybody in this company. We asked the questions to all the different people and no one could give us a definite, here's what you work off, this this this. So what we did, we went to [an SPC trainer] and said, "here is all our data for this year."

This quote also illustrates the *division of labour* and *knowledge* that is omnipresent in companies, and also a relevant feature of how people inhabit a space of reasons. Nobody knows everything that is relevant to producing a car. Each person is aware of a part, and his or her awareness is layered: Some reasons (in the wide sense) may be known well whereas others might be known in fragmented ways, only implicitly, or not at all.

Another example of the division of labour relates to the "assigned person" (AP) above. As soon as Kevin "raised an AP" he had done his job, and the problem was not his responsibility anymore. A third example is that Kevin and his colleagues fill in their charts, but know that it is the job of the SPC department (who are responsible for training and technical support) to calculate the mean and control limits from their data. This example illustrates that employees need not know all the statistical reasons behind SPC but do need to know something about the division of labour itself, and some of the statistical reasons to be able to communicate with others (team members, managers, suppliers) about their data and correctly fill in and interpret the control charts.

We were curious to what extent the SPC chart made sense to Kevin. When one of us (Res.) asked if the sideways histogram (to the right in Figure 1) helped him, he said:

> K.: It does and it doesn't because it just gives us an idea of where we are working. I mean that [histogram] is just little boxes to colour in for me [he laughs].
>
> Res.: It's not just supposed to be for little boxes to colour in. What could it tell you?
>
> K.: It tells you where about you are working. If you are running high or running low, but I concentrate on this [time series to the left] more because this tells me more than that [histogram]. That [histogram] will just give us an idea. Obviously you are supposed to get the peak in the middle behind the average line but that will tell us if we are running on average. I can look at the chart but to look at the histogram, I mean the obvious reason is that it will tell us if we are running just above the average or just below the average. (…) Really that should be in a nice spike right in the middle, right down the average line [he is probably describing the expected bell shape with the mode near the target line].

From such episodes we concluded that Kevin had a functional understanding of average (in relation to a target), variation (should be within certain limits) and distribution (roughly bell shaped) in relation to the mechanisms underlying the process. Such concepts are core in understanding and applying SPC. However, he kept calling control limits "spec limits," a phenomenon we have observed many times (Hoyles et al., 2007). As stated before, understanding control limits requires some understanding of standard deviations and the basics of the probabilistic rules in relation to the normal distribution. Crucially, despite their name, control limits are derived directly from the data, whereas spec limits are imposed externally. The trainers told us that this lack of understanding sometimes caused problems in communication between different groups in the company.

The second line in the quote above ("boxes to colour in") hints at a culture in which employees have to do things but not always know why. In fact, Kevin told us: "If you ask too many questions you end up doing a deep dive issue yourself, so really you are better just dropping a couple of slight hints and letting everybody else argue over it." The picture that emerged from such interviews is that employees tend to be aware of only that small part of the space of reasons that is directly relevant to their involvement in producing cars. However, to solve non-standard problems awareness of a larger part—including statistical reasons—is required, especially during nightshifts when few engineers or managers are around.

As another illustration of the importance of knowledge and how it is divided or distributed, we mention one finding from the interviews with three participants in an SPC course. We were surprised that the trainer asked participants to estimate standard deviations and calculate control limits by hand. Our impression was that they had only a limited idea what they were doing, and were actually hindered by the calculations rather than helped in their understanding. All participants we interviewed, however, appreciated having done the calculations once, just to know that these were done by the SPC department. Where our notion of understanding was focused on the statistical concepts involved in calculating the control limits, their take on understanding was *knowing about how labour and knowledge were divided*. Apart from being happy that these calculations were not part of their own work, they were also satisfied to note that the limits were not "conjured up" by management, but calculated on *their* data. In other words, we realised we had to enhance our notion of "understanding SPC" to a much wider notion in which the ways knowledge is distributed is taken into account. This implies that it is useful for employees to know when statistical inferences are made by others, and who these others are.

4. REFLECTIONS

To characterise statistical inference at work—the goal of this paper—we compare SPC with a form of statistical inference that is more widely known in education, hypothesis testing (4.1), and characterise the space of reasons in which SPC takes place more generally (4.2).

4.1. A COMPARISON OF HYPOTHESIS TESTING AND SPC

As we show below, the logic of SPC theory resembles that of hypothesis testing in some ways, but also differs from it in others. We start with the similarities to identify some candidate characteristics of statistical inference more generally.

1. In both types of statistical inference, the construct of interest has to be *measured*. In the airtightness example, the construct was air pressure in the car when blowing in air at a certain volume per time unit; this pressure was used as a measure of airtightness. *Samples* are used to predict some feature of a population or process.
2. Both approaches aim to detect *differences*, for example between hypothetical (expected, targeted) and the real measures of the population or process. In SPC the key idea is to detect trends in processes such as shifting means before measurements exceed specification limits.
3. The equivalent of a *null hypothesis* in SPC is “the process is stable” (there is no “significant” difference between the targeted measures and the real). This means that there is only common-cause variation: The mean and variability do not change much. The *alternative hypothesis* would then be “there is a change in the process” with a special cause, leaving unspecified what this cause might be. In this sense SPC shows more similarities to Fisher’s view on hypothesis testing than that of Pearson and Neyman, because Fisher did not require alternative hypotheses to be specified in advance (for the different views on hypothesis testing see for example Batanero, 2000; Biehler, 1982; Christensen, 2005).
4. In SPC, *probability-based rules* are used such as “seven points on either side of the mean may point to a special cause.” The choice of 7 seems to be based on practical effectiveness rather than theoretical significance of the 1/64 probability.
5. Possible *errors* in SPC resemble type I and II errors: Non-conforming points might be due to chance, and special causes might still not be detected by probability-based rules. In the airtightness example, Kevin noted leaks even when the measurements did not give reasons to think so.

These five points illustrate that the theory of SPC has an inferential structure that is in some respects similar to that of hypothesis testing. When comparing with Makar and Rubin’s (2007) three key characteristics of statistical inference, we can observe that all three are covered. Point 1 covers the issue of generalisation beyond the immediate data; data as evidence is a key point in both SPC and hypothesis testing; and points 4 and 5 hint at a probabilistic language.

The question arises whether the other similarities add anything specific or new. The first point on measurement is characteristic of statistical investigation in general, not only of statistical inference, and can therefore be excluded as a key feature of statistical inference. Points 2 and 3 are more interesting, because they point to something that is relevant to statistical inference but not explicitly covered in Makar and Rubin's list: The comparison of data with a model. In hypothesis testing we mostly compare the data measures with those of a hypothetical distribution; in SPC we compare data measures with those of a targeted distribution. What is underlying this is the view that data can be seen as model plus residuals or signal plus noise (Konold & Pollatsek, 2002).

Apart from the aforementioned similarities, there are also differences:

1. In the commonly used logic of hypothesis testing, the goal is to reject the null hypothesis. In SPC, however, the null hypothesis (the process being stable) is the desirable situation.
2. Hypothesis testing is focused on a single measurement of a statistic that is compared with the sampling distribution of that statistic. SPC, however, has a time dimension that is not typical of hypothesis testing. In fact, SPC could be seen as involving a series of different tests. In loose terms, we can say that SPC is focused on generalisations about a *process* rather than a population, though technically it is of course possible to frame a process as generating a population of measurements, many of which are still to come.
3. The focus of SPC is on the need for action, and the goal is to monitor special causes rather than to estimate the probability of a conditional statement which is the end result of hypothesis testing.
4. Hypothesis testing is mostly a formalised form of inductive inference. But in SPC the crucial inference to be made is to detect the special cause, which is in fact a form of abductive inference (finding an explanation for an anomaly in the data), which cannot be formalised.
5. In hypothesis testing one temporarily suppresses contextual information, whereas with SPC employees constantly use their contextual knowledge of the process to interpret data points. There may be perfectly good reasons to let the data be off target. For example, changing the process might be too expensive and not necessary for quality. In hypothesis testing, context should only be attended to before and after carrying out the statistical test, but not during the test. This means that the use of contextual information is much more "disciplined" (a term introduced by Pratt at the SRTL-5 conference) in hypothesis testing than in SPC.

This illustrates how the two types of inference are subject to different norms. Hypothesis testing is supposed to be independent of specific features of the situation, and contextual "noise" is left out during the calculations, whereas SPC is *pragmatic*. Hypothesis testing has become standardised whereas SPC is used in loose ways and often in non-standard ways. For instance, we have observed SPC use that we had not anticipated based on the SPC literature; for example, the use of control charts particularly when processes were unstable, whereas the literature typically recommends using SPC for stable processes (e.g., Oakland, 2003). Of course, hypothesis testing is sometimes used in loose or non-standard ways too, for example if conditions do not apply or if p-values do not tell what people really want to know (cf., Abelson, 1995).

4.2. SPACE OF REASONS AT WORK

To further characterise statistical inference at work, we address the space of reasons involved, in particular in SPC. Let us mention a few features we think are relevant.

1. A space of reasons encompasses what can be analytically distinguished as *contextual* and *statistical* reasons. The key issue however is that a holistic view on such reasons does not prioritise any of them *a priori*. Contextual or statistical reasons are

prioritised depending on whatever is required to reach a goal, such as delivering cars that are airtight enough and not too airtight.

In workplace statistical inference *contextual* reasons can be put into the foreground where statistical reasons might point in a different direction. In the case of the sealing material leading to airtightness being off-target, contextual constraints, particularly cost of implementation balanced against resulting productivity gains, led Kevin and his colleagues not to re-centre the process to its target. Such constraints emphasise the significance of one reason over another and hence form the background (part of the space of reasons) constituting attentiveness to one concern over another and hence to the taking of certain actions rather than others.

2. A space of reasons includes reasons informing both *statements (claims, judgements, etc.)* and *actions (decisions etc.)*. Statistical tests in educational settings are mostly focused on testing the veracity of statements. Employees, as the SPC examples illustrate, are often more concerned with the right action. Of course, better knowledge can lead to better decisions, but their focus is on meeting targets and being (more) efficient and productive, and it is these goals that constitute the normative background in which one concern figures as dominant and demanding of attention over another.

3. A space of reasons can be analysed at both a collective and an individual level. At an individual level we can focus on the reasons, implications, causes, and effects that a person is responsive to. Being responsive to reasons does not entail full awareness but only that judgements are made within such a space of reasons. At the collective level, we can envision the space of reasons as being constitutive of the community, practice, activity system, or context in which the inferences are made.

4. A space of reasons is necessarily normative. Some forms of drawing conclusions are culturally more acceptable than others. If someone makes a statement, we expect him or her to believe it and to able to give evidence for it. In scientific research (whether or not using formal statistical inference) the norm of credibility or even truth is important. But in a company the most important norm might be quality of the products or services defined in relation to efficiency and appeal to customers. Such norms have a major impact on what counts as evidence, inference, or conclusion.

So what have we gained by focusing on spaces of reasons instead of context, for example? First of all, we think that by starting with a space of reasons, we can temporarily overcome the distinction between statistics and context that is so deeply engrained in our thinking. This is particularly beneficial in situations in which it is no longer clear what is statistical or contextual. For example, in some plants of the factory, SPC has become part of the shopfloor "context," and in ideal cases, employees seamlessly coordinate what can be analytically separated as statistical and contextual reasons.

Second, when studying statistical inference it makes sense to focus on reasoning, and by highlighting "reasons," which we use as short for premises, conclusions, evidence, motives, purpose, utility, our knowledge of causes and effects, and so forth, we bring something to the fore that might be ignored if we strictly interpret "context" as location or setting.

5. DISCUSSION

5.1. A FEW OBSERVATIONS

This paper's goal was to characterise statistical inference at work and point to a type of statistical inference typically neglected in school-based educational research. We did so by comparing statistical process control as a commonly used technique for process improvement in industry with hypothesis testing. Next, we analysed the wider space of reasons in which statistical inference takes place. It turned out that there are commonalities that point to characteristics of statistical inference more generally. In addition to Makar and Rubin's (2007) three characteristics—generalisation, data as evidence, and probabilistic language—we also specified a fourth characteristic: comparing data with models, in particular, measures of the data with measures of a hypothetical or targeted distribution.

There are also differences. Firstly, SPC is pragmatic and focused on action. Secondly, there is a clearer place for abduction, whereas formal statistical inference tends to focus on induction. A third observation from our research in several companies is that inferences are mostly about *processes* rather than populations (see also Frick, 1998). In manufacturing sectors, data are mostly monitored to ensure that all items produced are within specification and to improve current production processes. When items are non-conforming or even fall out of specification, should the production process be stopped or is it more sensible to keep going? Stopping a car production line costs thousands of pounds per minute! In this setting, inferences are generally not based on formal statistical tests, but involve both statistical and contextual reasoning with a clear goal: maintaining the production process at a required accuracy or efficiency, or improving it.

The airtightness example emphasises *constraints of available resources* and the importance of the *division of labour and knowledge*. One advantage of using the notion of a space of reasons is that it helps to see human judgement as involving all such reasons including those that are beyond the more visible formal results of applying a statistical technique. Judging the consequences of having a batch of sealing material leading to a slightly different air pressure is one such example.

The adjective "informal" that is sometimes used in front of "statistical inference" does not seem to be suitable to characterise statistical inference at work. The theory of SPC is in fact formal to a certain extent: There are many books on the market on how to control production processes according to this theory. Hence, many people around the globe use SPC in somewhat similar ways. In that sense, SPC is more formal than, say, many intuitive forms of reasoning that young students display when they draw their first conclusions from data.

It could well be argued that what we emphasise as characteristics of the space of reasons in which statistical inference at work takes place applies to formal statistical inference as well. Indeed, we actually think that formal inferential techniques are used in such a way that we tend to forget that these techniques are used within a certain scientific space of reasons with particular norms and purposes, and that contextual knowledge is actually highly important in interpreting results of formal inference – despite the air of independence of contextual specificities that typically comes with significance tests. Moreover, we also face *division of labour* when using formal statistical inference: Even when we write research questions ourselves, collect the data ourselves, we use statistical software that others have programmed and most likely use statistical techniques that others have developed, before we arrive at our research conclusions. Hence we think that our analysis of statistical inference at work may shed light on statistical inference more generally.

5.2. LIMITATIONS

With the choice of SPC as a prototypical example of statistical inference commonly used in industry, we have restricted ourselves to one type of statistical inference. It is perfectly possible that the analysis of other statistical techniques, perhaps in other workplaces, would lead to additional or other characteristics.

Another limitation of our research seems inherent to workplace studies: As outsiders it is extremely hard to gain access to companies: Time is money. Hence the companies we have studied form a convenience sample, and our time with employees, whether in interviews or training courses, was short compared to much research in school settings. This explains why we have not been able to do any workshadowing to the extent that we could observe employees actually solve problems by using SPC charts. Nor were we allowed to videotape employees or ask them to take tests to identify their level of statistical understanding. We were confined to stories told by managers, operators, and engineers, which meant that we had little opportunity to study inference in action and that we were not able to draw many generalised conclusions.

5.3. IMPLICATIONS FOR FUTURE RESEARCH AND EDUCATION

The aforementioned limitations naturally lead to implications for future research: It would be interesting to study more types of statistical inference in several workplaces, preferably in action. Another recommendation for future research is to characterise more generally what employees exactly need to know in different sectors of work. It is, for instance, hard to pinpoint to which level of formality, and in what sense, employees need to understand statistical concepts. Of course, they need to understand what the sources of variation are and what variation looks like in graphs (cf., Noss et al., 2007; Wild & Pfannkuch, 1999); they need to reason with a notion of distribution, mean versus target, spread, and measures of spread; they need to be able to interpret graphs, and so forth. Working with machines, employees need to know about causes and effects, but also how independent variables influence a dependent one. In short, employees need to know the key aspects of the *model* at stake, that is, the relationships between relevant variables and the causes and effects of changing those variables. The model need not be purely statistical or mathematical (Bakker, Hoyles, Kent, & Noss, 2006): Kevin's model of the variables of air volume and pressure were related to airtightness of cars was context specific. Yet we suggest adding the understanding of such a—possibly situated—model to Rubin et al.'s (2006) list of statistical ideas underpinning statistical inference.

Despite the advocated integration of statistical and contextual knowledge in practice, the statistics courses that we have observed and heard about in industry generally provide participants with little opportunity to connect what they have learned about statistics to their practice. Newly acquired statistical knowledge thus often stays separated from the rich contextual knowledge employees have of their work processes, instead of being perceived as an organic part of a space of reasons.

A more theoretical implication for research is the need to explore the consequences of framing educational research in terms of spaces of reasons and the tradition of inferentialism from which such philosophical constructs stem. They may provide us with a useful perspective on training in workplaces. Instead of primarily asking ourselves which statistical theory employees need to learn, we should perhaps ask, "How can employees become adept in their workplace's space of reasons? How can the reasons they attend to be enhanced by knowledge of quality improvement strategies?" In another paper we report on how we have tried this in our first design experiments (Hoyles et al., 2007).

Formulating potential implications of workplace research for school education is a tricky business. As Säljö (2003) notes, one should not make the mistake to try and copy workplace situations in school education. School is a different system with different goals than workplaces. Nor should we necessarily adapt our language: School and workplace situations are constituted by very different spaces of reasons. The following implications therefore have to be interpreted only as a tentative list for discussion purposes:

1. Not only in workplace training but also in school settings, we should acknowledge the importance of context knowledge, real-world constraints, actions, and responsibility, and not confine the theory and learning activities to clean noise-free examples. What is required, also when teaching hypothesis testing, is to emphasise that drawing on context knowledge is disciplined. This might mean that courses that introduce formal statistical inference to students ideally also spend time afterwards on how formal statistical tests are actually used in practice, within a wider space of reasons.

2. We should pay attention to the mechanisms that cause variation, because then variation becomes easier for students to understand (cf., Wild & Pfannkuch, 1999). Yet there is a need to generalise and become familiar with statistical measures that can be applied in many other situations.

3. In our view, school-based educational research should pay attention to student understanding of processes in addition to populations because our anecdotal evidence suggests that many employees will deal with processes and not just with populations. The production of widgets at school (e.g., Konold & Lehrer, in press) might be a useful context to learn about many relevant statistical ideas.

ACKNOWLEDGEMENTS

This paper is based on an earlier version presented at the fifth research forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5), 11-17 August 2007, Warwick, UK. We thank all participants for fruitful discussions about informal statistical inference, and Iddo Gal, Cliff Konold, and Rolf Biehler for their feedback on earlier versions of this paper. Kees Klaassen and Bart Ormel have also made very helpful suggestions. We further gratefully acknowledge funding by the UK Economic and Social Research Council's Teaching and Learning Research Programme, Grant number L139-25-0119.

REFERENCES

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bakker, A., Derry, J., & Konold, C. (2006). Technology to support diagrammatic reasoning about center and variation. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D4_BAKK.pdf]

Bakker, A., Hoyles, C., Kent, P., & Noss, R. (2006). Improving work processes by making the invisible visible. *Journal of Education and Work, 19*, 343-361.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning*, *2,* 75-97

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D1_BENZ.pdf].

Biehler, R. (1982). *Explorative Datenanalyse – Eine Untersuchung aus der Perspective einer deskriptiv-empirischen Wissenschaftstheorie* [Exploratory data analysis - An investigation from the perspective of a descriptive empirical scientific theory] Bielefeld, Germany: Universität Bielefeld.

Caulcutt, R. (1995). The rights and wrongs of control charts. *Applied Statistics, 44*, 279-288.

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician, 59*, 121-126.

Cobb, G. W. (1999). Discussion of "Let's use CQI in our statistics programs." *The American Statistician, 53*, 16-21.

Frick, R. W. (1998). Interpreting statistical testing: Process and propensity, not population and random sampling. *Behavior Research Methods, Instruments, and Computers, 30*, 527-535.

Hoyles, C., Bakker, A., Kent, P., & Noss, R. (2007). Attributing meanings to representations of data: The case of statistical process control. *Mathematical Thinking and Learning, 9,* 331-360.

Konold, C., & Lehrer, R. (in press). Technology and mathematics education: An essay in honor of Jim Kaput. In L. D. English (Ed.). *Handbook of international research in mathematics education* (2nd ed.). New York: Routledge.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33,* 259-289.

Lave, J. (1988). *Cognition in practice: Mind, mathematics and culture in everyday life*. Cambridge, UK: Cambridge University Press.

McDowell, J. (1996). *Mind and world* (2nd ed.). Cambridge, MA: Harvard University Press.

Makar, K., & Rubin, A. (2007, August). *Beyond the bar graph: Teaching informal statistical inference in primary school.* Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), University of Warwick, UK.

Noss, R., Bakker, A., Hoyles, C., & Kent, P. (2007). Situating graphs as workplace knowledge. *Educational Studies in Mathematics, 65,* 367-384.

Noss, R., & Hoyles, C. (1996). *Windows on mathematical meanings: Learning cultures and computers.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Noss, R., Pozzi, S., & Hoyles, C. (1999). Touching epistemologies: Meanings of average and variation in nursing practice. *Educational Studies in Mathematics, 40,* 25-51.

Nunes, T. A., Schliemann, D., & Carraher, D. W. (1993). *Street mathematics and school mathematics*. Cambridge, UK: Cambridge University Press.

Oakland, J. S. (2003). *Statistical process control* (5th ed.). Amsterdam: Butterworth-Heinemann.

Paparistodemou, E., & Meletiou-Mavrotheris, M. (2007, August). *Enhancing reasoning about statistical inference in 8 year-old students.* Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), University of Warwick, UK.

Pfannkuch, M. (2006). Informal inferential reasoning. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/6A2_PFAN.pdf]

Rubin, A., Hammerman, J. K. L., & Konold, C. (2006). Exploring informal inference with interactive visualization software In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International*

Conference on Teaching Statistics, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D3_RUBI.pdf]

Säljö, R. (2003). Epilogue: From transfer to boundary-crossing. In T. Tuomi-Gröhn & Y. Engeström (Eds.), *Between school and work: New perspectives on transfer and boundary-crossing* (pp. 311-321). Amsterdam: Elsevier.

Smith, J. P. (1999). Tracking the mathematics of automobile production: Are schools failing to prepare students for work? *American Educational Research Journal, 36,* 835-878.

Steen, L. A. (2003). Data, shapes, symbols: Achieving balance in school mathematics. In B. L. Madison & L. A. Steen (Eds.), *Quantitative literacy: Why literacy matters for schools and colleges* (pp. 53-74). Washington, DC: The Mathematical Assiociation of America. Retrieved October 27, 2007 from http://www.maa.org/ql/qltoc.html

Techno-mathematical Literacies Project (2003-2007). www.lkl.ac.uk/technomaths

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistics Review, 67,* 223-248.

Zieffler, A., Garfield, J., & delMas, R. (2007, August). *Studying the development of college students' informal reasoning about statistical inference.* Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), University of Warwick, UK.

ARTHUR BAKKER
University of London
London Knowledge Lab
23-29 Emerald Street
London WC1N 3QS
United Kingdom

Currently working at:
Utrecht University
Freudenthal Institute
PO Box 9432
3506 GK Utrecht
The Netherlands

PAST IASE CONFERENCES

JOINT ICMI /IASE STUDY
STATISTICS EDUCATION IN SCHOOL MATHEMATICS: CHALLENGES FOR TEACHING AND TEACHER EDUCATION
Monterrey, Mexico, June 30 - July 4, 2008

The International Commission on Mathematical Instruction (ICMI, http://www.mathunion.org/ICMI/) and the International Association for Statistical Education (IASE, http://www.stat.auckland.ac.nz/~iase/) organised the Joint ICMI /IASE Study Statistics Education in School Mathematics: Challenges for Teaching and Teacher Education.

Accepted papers were presented in the Conference and appeared in the Proceedings that were published by ICMI and IASE as a CD-ROM and on the Internet. The second part of the Joint Study—the Joint Study book—was produced after the conference and is published in the ICMI Study Series.

More information: Carmen Batanero, batanero@ugr.es

Website with links to proceedings and Joint Study book:
http://www.ugr.es/~icmi/iase_study/

ICME 11
INTERNATIONAL CONGRESS ON MATHEMATICAL EDUCATION
TOPIC STUDY GROUP # 13
RESEARCH AND DEVELOPMENT IN THE TEACHING AND LEARNING OF PROBABILITY
Monterrey, Mexico, July 6 - 13, 2008

Probability and statistics education are relatively new disciplines. Both have only recently been introduced into the main stream school curricula in many countries. While the application oriented statistics is undisputed in its relevance, discussion about probability is more ambivalent. Reduction of probability to the classical conception, mainly based on combinatorics, or its tight connection to higher mathematics, is an argument to abandon this part of the discipline in favour of the statistics part. However, there are some arguments for a strong role for probability within stochastics curricula:

1. Misconceptions on probability affect people's decision in important situations, such as medical tests, jury verdict, investment, assessment, etc.
2. Probability is essential to understand any inferential procedure of statistics.
3. Probability offers a tool for modelling and "creating" reality. For example, modern physics cannot be formulated without reference to probability concepts. The concepts of risk (not only at financial markets) and reliability are closely related to and dependent upon probability.

Thus the challenge is to teach probability in order to let the students understand it. The focus has to be on creating approaches to probability that are more accessible and motivating. Additionally, the frequentist and subjectivist views of probability, and connections of probability to practical applications should be taken into account.

Statistics Education Research Journal, 7(2), 146-156, http://www.stat.auckland.ac.nz/serj
© International Association for Statistical Education (IASE/ISI), November, 2008

Simulation is one such strategy, as is visualization of abstract concepts; there are more. The use of technology also enables to reduce the calculation technicalities and focus the learner on the concepts instead. The world of personal attitudes and intuitions is another source for success or failure of teaching probability. With these challenges in mind, we have encouraged in our call papers and presentations related to the following topics:

- Individuals' corner
  Students' understanding and misunderstanding of fundamental probabilistic concepts
  Ideas of probability in young children
- Impact of technology
  The use of technology for students' learning of probability
  Using specific software (Fathom, probability explorer, etc.) to study probability
  Special issues in e-learning
- Teacher's corner
  Teacher education on the topic of probability
  Teachers' conceptions about teaching probability
- Fundamental ideas
  The probabilistic idea of a random variable; distribution and expectation
  The central limit theorem; convergence
  Bayes' theorem and conditional probability; independence; exchangeability
  Probabilistic modelling – a probabilistic look at distributions

Out of the papers submitted, 17 papers were accepted after careful examination. We grouped them to the following topics which were the themes of our sessions at the conference:

- Issues in Probability Teaching and Learning,
- Informal Conceptions,
- Conditional probability and Bayes' theorem.

A panel discussion on the topic "Fundamental Ideas in Probability Teaching at School Level" completed our programme. The panel focused on the impact of recent trends in school curricula, which have removed probability at early stages in favour of data analysis techniques.

The authors came from Europe, USA, Australia and Latin America, the English and the Spanish worlds, and the "rest" were distributed "evenly." More details of the presentations are available from the conference website (see below), including full papers and PowerPOint shows as well. A critical review will follow.

The hope is that ICME will continue to organize topic study groups on probability and statistics separately. With this "strategy" we did in fact split our potential audience as all the study groups are held at the same time slots. However, the great interest in our group on probability as well as the number of people who attended the parallel statistics group confirm that we can attract many more people to our topic by two separate groups. The split into the two groups also allowed for a more convenient focus of the pertinent presentations and discussions. It showed that—against the international trend towards statistics and away from probability in all international curricula—there is still a substantial interest in research in probability issues as it is highly relevant for any teaching and learning of statistics. This holds also for the joint study of ICME and IASE, which was held one week prior to the ICME congress where a panel discussion about a vital role for probability within curricula met a strong echo and led to a lively discussion on the role of probability within educational research and in curricula.

To view the presentented papers and other activities available on the ICMI website:
http://tsg.icme11.org/tsg/show/14

ICME 11
INTERNATIONAL CONGRESS ON MATHEMATICAL EDUCATION
TOPIC STUDY GROUP # 14
RESEARCH AND DEVELOPMENT IN THE TEACHING AND LEARNING OF STATISTICS
Monterrey, Mexico, July 6 - 13, 2008

Statistics education is a growing field of research and development at school and university level. The topic group focused on presenting and discussing recent research.
Statistics at school level is usually taught in the mathematics classroom in connection with learning probability. Inferential statistics is based on basic understandings of probability. Our topic includes probabilistic aspects in learning statistics, whereas research with a specific focus on learning probability is being discussed TSG 13 of ICME.

To view the presentented papers and other activities are available on the ICMI website: http://tsg.icme11.org/tsg/show/15

OTHER PAST CONFERENCES

6TH AUSTRALIAN CONFERENCE ON TEACHING STATISTICS
Melbourne, Australia, July 3 - 4, 2008

The 6th OZCOTS was held as a satellite to the Australian Statistical Conference. Presented invited and contributed papers and forums on topics across the tertiary statistical education spectrum are of interest to statisticians, statistical educators, and the statistical profession. OZCOTS 2008 and its invited speakers are associated with a National Senior Teaching Fellowship on the teaching and assessment of statistical thinking within and across disciplines.

More information and link to proceedings:
http://silmaril.math.sci.qut.edu.au/ozcots2008/

CensusAtSchool: 2ND INTERNATIONAL WORKSHOP
Los Angeles, California, July 28 - 29, 2008

The International CensusAtSchool project encourages the use of real data, from and about school children, and promotes the teaching and learning of statistical thinking skills in the classroom. This gives children increased understanding of data, wherever it originates, and encourages them to develop a healthy skepticism towards statistics that are constantly presented to them by the media and the society they live in.

More information: Juana Sanchez (jsanchez@stat.ucla.edu)
Website with links to presentations and other materials:
http://censusatschool-california.stat.ucla.edu

2008 JOINT STATISTICAL MEETINGS
Denver, CO, USA, August 3 - 7, 2008

JSM (the Joint Statistical Meetings) is the largest gathering of statisticians held in North America. It is held jointly with the American Statistical Association, the International Biometric Society (ENAR and WNAR), the Institute of Mathematical Statistics, and the Statistical Society of Canada.

Website: http://www.amstat.org/meetings/jsm/2008/

FORTHCOMING IASE CONFERENCES

ISI-57
THE 2009 SESSION OF THE INTERNATIONAL STATISTICAL INSTITUTE
Durban, South Africa, August 16 – 22, 2009

IASE sponsored Invited Paper Meetings for 57th Session in Durban are being organised by Helen MacGillivray (Australia, h.macgillivray@qut.edu.au). The IASE Programme Committee for ISI-57 has chosen the theme "Statistics Education for the Future."

IASE has nine IPM (Invited Paper Meeting) sessions, two of which include issues raised by the local organisers, and has two joint sessions with IAOS.

Session Number	Section representation	Title of Invited Paper Meeting	Organiser(s)
IPM15	IAOS IASE	The challenge of building a supply of statisticians for the future	To be determined, c/o Nancy McBeth, Nancy.McBeth@stats.govt.nz
IPM36	IASE IAOS	The roles of statistical agencies in developing statistical literacy	Reija Helenius, Finland, Reija.Helenius@stat.fi
IPM37	IASE Local Hosts	Educating the public on how to use official statistics	Peter Wingfield-Digby, pwdigby@loxinfo.co.th
IPM38	IASE Local Hosts	Challenges faced in Statistics Education in African countries	Delia North, South Africa, northd@ukzn.ac.za
IPM39	IASE	Balancing the training of future statisticians for workplace and research	Charles Rohde, USA, crohde@jhsph.edu
IPM40	IASE	Exploiting the progress in statistical graphics and statistical computing for the benefit of statistical literacy	Juana Sanchez, USA, jsanchez@stat.ucla.edu
IPM41	IASE	Survey research in statistics education	Irena Ograjensek, Slovenia, irena.ograjensek@ef.uni-lj.si
IPM42	IASE	Research on informal inferential reasoning	Katie Makar, Australia, k.makar@uq.edu.au
IPM43	IASE	Teaching, learning and assessing statistics problem solving in higher education	Neville Davies, UK, neville.davies@ntu.ac.uk
IPM44	IASE	Technologies for learning and teaching in developing countries	Gabriella Belli, USA, gbelli@vt.edu
IPM45	IASE	Virtual learning environments for statistics education	Adriana Backx Noronha Viana, Brazil, backx@usp.br and Pieternel Verhoeven, Netherlands, n.verhoeven@roac.nl

The website http://www.statssa.gov.za/isi2009/ has information on all matters relating to ISI 2009, including important dates, and will be regularly updated as new information develops.

More information: Helen MacGillivray, h.macgillivray@qut.edu.au

2009 IASE SATELLITE CONFERENCE TO THE 57TH SESSION OF THE ISI "NEXT STEPS IN STATISTICS EDUCATION" Durban , South Africa, August 14 -15, 2009 (Immediately before ISI 57 in Durban)

All submissions addressing the theme "Next Steps in Statistics Education" will be welcome. This theme has been chosen to particularly attract papers under the following headings:

1. What constitutes best practice for the curriculum beyond the "Introductory Statistics" course? What courses should follow on for those wishing to major in Statistics and what additional training should we offer to those in other disciplines?
2. What elements of our undergraduate curriculum specifically prepare our students for their careers post-graduation, either in the workplace or as masters/doctoral students? How can we improve these elements?
3. Now that more countries have school curricula that include substantial emphasis on data and chance, how can we better prepare teachers for implementing those curricula? What curricular materials and tools can we develop to improve students' learning of statistics at school level?
4. Since the 1949 formation of its precursor, the ISI Statistical Education Committee, the IASE has matured as an organisation. As we move towards ICOTS 8, we note that great progress has already been made in the field of Statistics Education but the challenge we face now is to consider the next steps that we must take. How can we build on past progress to raise the profile of our field so that it becomes a more visible and vibrant pursuit?

More information can be found on conference webpage:
http://www.ucd.ie/statdept/2009_iase_satellite.html
Conference email: IASE_Satellite@maths.ucd.ie

SRTL-6 THE SIXTH INTERNATIONAL RESEARCH FORUM ON STATISTICAL REASONING, THINKING, AND LITERACY The Role of Context and Evidence in Informal Inferential Reasoning Brisbane, Australia, July 10 - 16, 2009

The sixth in a series of International Research Forums on Statistical Reasoning, Thinking and Literacy (SRTL-6) is to be held in Brisbane, Australia from July 10 to July 16, 2009. The School of Education at The University of Queensland, will host the Forum. The Forum's focus will build on the work presented and discussed at SRTL-5 on informal ideas of statistical inference. Recent research suggests an important role for developing ideas of informal types of statistical inference even at early educational levels. Researchers have developed instructional activities that encourage students to infer beyond samples of data and use technological tools to support these informal inferences.

The findings of these studies reveal that the context of the data and the use of evidence may be important factors to study further. The role of context is of particular interest because in drawing (informal) inferences from data, “students must learn to walk two fine lines. First, they must maintain a view of data as 'numbers with a context'” (Moore, 1992). At the same time, “they must learn to see the data as separate in many ways from the real-world event they observed” (Konold & Higgins, 2003, p. 195). That is, they must abstract the data from that context. The role of evidence is also of particular interest because in learning how to make data-based claims (argumentation), students must consider the evidence used to support the claim, the quality and justification of the evidence, limitations of the evidence and finally, an indication of how convincing the argument is (Ben-Zvi, Gil, & Apel, 2007).

Based on SRTL-5, we characterize Informal Inferential Reasoning (IIR) as the cognitive activities involved in drawing conclusions with some degree of uncertainty that go beyond the data and having empirical evidence for them. Three principles appear to be essential to informal inference: (1) generalizations (including predictions, parameter estimates, and conclusions) that go beyond describing the given data; (2) the use of data as evidence for those generalizations; and (3) conclusions that express a degree of uncertainty, whether or not quantified, accounting for the variability or uncertainty that is unavoidable when generalizing beyond the immediate data to a population or a process (Makar & Rubin, 2007).

An interesting range of diverse research presentations and discussions have been planned and we look forward to a stimulating and enriching gathering. These papers will address the role of context and evidence when reasoning about informal inference at all levels of education including the professional development of elementary and secondary teachers.

The structure of the scientific program will be a mixture of formal and informal sessions, small group and whole group discussions, and the opportunity for extensive analysis of video-taped research data. There will also be a poster session for exhibiting current research of participants on additional topics related to statistics education. The Forum is co-chaired by Dani Ben-Zvi (University of Haifa, Israel) and Joan Garfield (University of Minnesota, USA), locally organized by Katie Makar and Michael Bulmer (The University of Queensland), and planned by a prestigious international advisory committee. Conference attendance is by invitation only.

For more information, visit the SRTL website at: http://srtl.stat.auckland.ac.nz/ or email SRTL2009@gmail.com.

ICOTS-8
DATA AND CONTEXT IN STATISTICS EDUCATION: TOWARDS AN EVIDENCE-BASED SOCIETY
Ljubljana, Slovenia, July 11-16, 2010

The 2010 International Conference on Teaching Statistics will be held in the city of Ljubljana, Slovenia, July 11-16. It is being organised by the IASE and the Slovenian Statistical Association. The venue will be the Ljubljana Cultural and Congress Centre.

Statistics educators, statisticians, teachers and educators at large are invited to contribute to the scientific programme. Types of contribution include invited papers, contributed papers and posters. No person may author more than one

Invited Paper at the conference, although the same person can be co-author of more than one paper, provided each paper is presented by a different person.

Voluntary refereeing procedures will be implemented for ICOTS-8. Details of how to prepare manuscripts, the refereeing process and final submission arrangements will be announced later.

INVITED PAPERS

Invited Paper Sessions are organized within 10 Conference Topics as follows.

Topics and Topic Convenors

1. Data and Context in Statistics Education: Towards an Evidence-based Society.
   Brian Phillips (Australia)   bphillips@swin.edu.au
   Irena Ograjensek (Slovenia)   irena.ograjensek@ef.uni-lj.si
2. Statistics Education at the School Level.
   Mike Shaughnessy (USA)   mikesh@pdx.edu
   Doreen Connor (UK)   doreen.connor@ntu.ac.uk
3. Learning to Teach Statistics.
   Katie Makar (Australia)   k.makar@uq.edu.au
   Joachim Engel (Germany)   engel@math.uni-hannover.de
4. Statistics Education at the Post Secondary Level.
   Elisabeth Svensson (Sweden)   elisabeth.svensson@esi.oru.se
   Larry Weldon (Canada)   weldon@sfu.ca
5. Assessment in Statistics Education.
   Beth Chance (USA)   bchance@calpoly.edu
   Iddo Gal (Israel)   iddo@research.haifa.ac.il
6. Statistics Education, Training and the Workplace
   Gabriella Belli (USA)   gbelli@vt.edu
   Peter Petocz (Australia)   peter.petocz@mq.edu.au
7. Statistics Education and the Wider Society
   Richard Gadsden (UK)   R.J.Gadsden@lboro.ac.uk
   Oded Meyer (USA) meyer@stat.cmu.edu
8. Research in Statistics Education
   Arthur Bakker (The Netherlands)   a.bakker@fi.uu.nl
   Tim Burgess (New Zealand) t.a.burgess@massey.ac.nz
9. Technology in Statistics Education
   Deborah Nolan (USA)   nolan@stat.berkeley.edu
   Paul Darius (Belgium)   paul.darius@biw.kuleuven.be
10. An International Perspective on Statistics Education
    Delia North (South Africa)   northd@ukzn.ac.za
    Enriqueta Reston (Phillipines)   edreston@usc.edu.ph

Session themes within each Topic are organized. The themes and Session organizers with email contact are available on the ICOTS-8 web site http://icots8.org/, under "Scientific Programme." The list of invited speakers is close to completion. A few gaps remain. If you are interested in being considered to fill one of these contact the Programme Chair John Harraway (jharraway@maths.otago.ac.nz) by December 31, 2008.

CONTRIBUTED PAPERS

Contributed paper sessions will be arranged in a variety of areas. Those interested in submitting a contributed paper should contact either Gilberte Schuyten (Gilberte.Schuyten@UGent.be), John McKenzie (mckenzie@babson.edu) or Flavia

Jolliffe (F.Jolliffe@kent.ac.uk) before August 31, 2009 if being refereed or before 30 November, 2009 if not being refereed.

POSTERS

Those interested in submitting a poster should contact Mojca Bavdaz (mojca.bavdaz@ef.uni-lj.si) or Alesa Lotric Dolinar (alesa.lotric.dolinar@ef.uni-lj.si) before January 15, 2010.

GENERAL ISSUES

More information is available from the ICOTS-8 web site at http://icots8.org/ which will continue to be updated over the next two years, or from the ICOTS IPC Chair John Harraway, (jharraway@maths.otago.ac.nz), the Programme Chair, Roxy Peck (rpeck@calpoly.edu) and the Scientific Co-ordinator, Helen MacGillivray (h.macgillivray@qut.edu.au).

OTHER FORTHCOMING CONFERENCES

USCOTS 2009
UNITED STATES CONFERENCE ON TEACHING STATISTICS
“LETTING GO TO GROW”
Columbus, OH, USA, June 25 - 27, 2009

The third biennial United States Conference on Teaching Statistics (USCOTS 09) will be held on June 25-27, 2009 at the Ohio State University in Columbus, Ohio, hosted by CAUSE, the Consortium for the Advancement of Undergraduate Statistics Education. The target audience for USCOTS is teachers of undergraduate and AP statistics, from any discipline or type of institution. Teachers from two-year colleges are particularly encouraged to attend.

The theme for USCOTS 2009 is Letting Go to Grow. “Letting Go” has many interpretations, such as letting go of some classic course content in order to better align with course goals, letting go of some old ideas about pedagogy in order to use more effective methods, or letting go of old notions about the students we teach in order to better facilitate their learning. USCOTS is a “working conference” with many opportunities for hands-on activities, demonstrations, networking, idea sharing, and receiving the latest information on research and best practices in teaching statistics. Leaders in statistics education and assessment will give plenary talks, including Dani Ben-Zvi (Haifa, Israel), George Cobb (USA), Peter Ewell (USA), Ronald Wasserstein (USA), and Chris Wild (Auckland, New Zealand).

Details are available at USCOTS web page: http://www.causeweb.org/uscots

10TH INTERNATIONAL CONFERENCE OF THE MATHEMATICS EDUCATION INTO THE 21ST CENTURY PROJECT
MODELS IN DEVELOPING MATHEMATICS EDUCATION
Dresden, Saxony, Germany, September 11 – 17, 2009

The Mathematics Education into the 21st Century Project was founded in 1986 and is dedicated to the planning, writing and disseminating of innovative ideas and materials in Mathematics and Statistics Education. You are warmly invited to attend our 10th anniversary conference in the heart of the historic city of Dresden, Germany. The conference is organized in full cooperation with the Saxony Ministry of Education. All our conferences have a strong Statistics Education component.

INTERNATIONAL ORGANISERS

Dr. Alan Rogerson, Coordinator of the Mathematics in Society Project (Poland)
Professor Fayez Mina, Faculty of Education, Ain Shams University (Egypt)

CHAIR OF THE LOCAL ORGANISING COMMITTEE

Prof. Dr. Ludwig Paditz, Dresden University of Applied Sciences.

Further information: Alan Rogerson, arogerson@inetia.pl
Web site: http://math.unipa.it/~grim/21project.htm

2009 JOINT STATISTICAL MEETINGS
Washington, DC, USA, August 1-6, 2009

JSM (the Joint Statistical Meetings) is the largest gathering of statisticians held in North America. It is held jointly with the American Statistical Association, the International Biometric Society (ENAR and WNAR), the Institute of Mathematical Statistics, and the Statistical Society of Canada. Attended by over 5000 people, activities of the meeting include oral presentations, panel sessions, poster presentations, continuing education courses, exhibit hall (with state-of-the-art statistical products and opportunities), career placement service, society and section business meetings, committee meetings, social activities, and networking opportunities.

More information: jsm@amstat.org

Website: http://www.amstat.org/meetings/jsm/2009/

STATISTICS EDUCATION RESEARCH JOURNAL REFEREES DECEMBER 2007-NOVEMBER 2008

Acknowledgement: We are grateful to the following who acted as referees for the *Statistics Education Research Journal* in the past year.

Janet Ainley, UK
Arthur Bakker, The Netherlands
Dani Ben Zvi, Israel
Monique Bijker, The Netherlands
Carol Joyce Blumberg, USA
Gail Burrill, USA
Matt Ciancetta, USA
Michelle Everson, USA
Paul Fields, USA
Joan Garfield, USA
Juan Godino, Spain
Randall Groth, USA
Peter Johnston-Wilder, UK
Jim Hammerman, USA
Flavia Jolliffe, UK
Sibel Kazak, USA
Katie Makar, Australia
Lionel Pereira Mendoza, Canada
Carlos Monteiro, Brazil
James Nicholson, UK
Jennifer Noll, USA
Ann O'Connell, USA
Elena Papanastasiou, Cyprus
Peter Petocz, Australia
Dave Pratt, UK
Chris Reading, Australia
Anna Reid, Australia
Andee Rubin, USA
Luis Saldanha, USA
Ernesto Sanchez, Mexico
Peter Sedlmeir, Germany
James Tarr, USA
Brandon Vaughn, USA
Jane Watson, Australia
Helena Wessels, South Africa
Ryan Zayac, USA
Andy Zieffler, USA