# On the improvement of students' conceptual understanding in statistics education

Luc Budé

Omslagontwerp: Frenk van Hooydonk

ISBN/EAN: 978-90-5681-262-1
NUR code: 860

# On the improvement of students' conceptual understanding in statistics education

Proefschrift

ter verkrijging van de graad van doctor
aan de universiteit Maastricht
op gezag van de rector magnificus
Prof. Mr. G. P. M. F. Mols
volgens het besluit van
het College van Decanen
in het openbaar te verdedigen

op vrijdag 5 oktober 2007
om 12.00 uur

door
Lucas Michel Budé

*Promotor*
Prof. Dr. M. P. F. Berger

*Copromotores:*
Dr. Tj. Imbos
Dr. M. W. J. van de Wiel

*Beoordelingscommissie*
Prof. Dr. C. P. M. van der Vleuten (voorzitter)
Dr. D. H. J. M. Dolmans
Prof. Dr. W. H. Gijselaers
Prof. Dr. F. G. W. C. Paas (Open Universiteit, Erasmus Universiteit)
Prof. Dr. G. Schuyten (Universiteit Gent)

Het onderzoek dat ten grondslag ligt aan deze dissertatie, is uitgevoerd binnen het project: ‘Onderwijs in moeilijk leerbare kennisdomeinen’. Dit project is mede gefinancierd uit het voormalig profileringfonds van de Universiteit Maastricht.

# Contents

# Chapter 1

## On the improvement of students' conceptual understanding in statistics education

## 1.1 Introduction

Students in social and health sciences often find the topic of statistics difficult, it is hard for them to understand (Garfield, 2003), and it appears that they dislike it (Gal, Ginsberg & Schau, 1997). So, there seems to be a need for innovations in statistics education directed at two points of interest: the improvement of students' learning and the stimulation of motivation. The aim of the studies in this dissertation is to investigate the effect of modifications in statistics education on students' learning and on a number of motivational components, and the influence of both on students' conceptual understanding of the subject matter.

### *1.1.1 Improving education*

Statistics education, like all forms of education, has two related aspects, namely teaching and learning. Teaching can be modified and adjusted by instructors. Ergo, instructors can try to design ways of teaching that improve student learning. The assumption is that improved learning will lead to improved conceptual understanding. Designing ways of teaching that directly influence student learning, involves attempts to make the subject matter clearer, more comprehensible and thus easier to learn for the student. Instructors also try to design ways of teaching that make it easier for students to apply what they have learned; i.e. ways of teaching that make it easier for the students to put their knowledge into practice. Such attempts that directly act upon learning are characterised by, for example, trying to select appropriate learning material, improving lectures, writing better text books, selecting and applying computer software, and upgrading practical training sessions.

Learning, however, involves more than that. Learning is also dependent on the effort students are willing to invest and on their style of studying, which both are strongly related to other motivational aspects. Motivation is a term that is often loosely used to refer to related constructs, such as attribution, affect, self-efficacy, goal orientation, perceived competence, and interest, without attempts at a more precise definition (Murphy & Alexander, 2000). However, despite the lack of a generally accepted definition of motivation, it can be conceptualised as the constellation of interrelated thoughts, feelings, and behaviour a person displays. For example, defeatist thoughts may be linked with negative affect which may be linked to passive behaviour (Peterson, Maier & Seligman, 1993; Pintrich & Schunk, 1996). Simply put, if a student has two learning tasks, and one of them is thought to be interesting, enjoyable, and easy; the other one boring, unpleasant, and difficult, then it is likely that the student will put more effort into the first learning task and study the corresponding material more thoroughly. This supposition seems trivial, yet from a review of the literature on educational research it can be concluded not much research has been done on ways of teaching that make learning more pleasant and enjoyable. Affective aspects are usually not highly valued.

Research in statistics education is not an exception in this respect. A broad search of the statistics education literature of the past fifteen years only yielded one article specifically about affect. It addresses humour in the classroom

(Friedman, Friedman & Amoo, 2002). The authors only give examples, descriptions, and assumed effects; they did not manipulate humour and measure the established effects. This is the same for articles on attitudes toward statistics. The attitude of a person also includes affective aspects (Gal et al., 1997). We only found a limited number of studies on attitudes toward statistics. All of these studies were directed at correlations between attitudes and some outcome variable (e.g., Johnson & Dasgupta, 2005; Mvududu, 2003; Vanhoof, Castro Sotos, Onghena, Verschaffel & Van Dooren, 2006). No experimental manipulations are reported in the literature or recommendations on how to improve attitudes, affect or other motivational aspects.

Studies in statistics education are mainly focussed at innovations of the learning environment and concrete methods and tools for instruction directly aimed at improving student learning and performance (Garfield, delMas & Chance, 1999). For example, delMas, Garfield, and Chance (1999) have shown that multimedia simulation activities improve students' statistical reasoning. West and Ogden (1998) give examples of technological innovations (Java applets) that stimulate student learning. Some studies aimed at the activation of the students. Ferrall (1995) presented interactive tutorials in stata, which were intended to improve comprehension. Steinhorst and Keeler (1995) studied the effect of what they called conceptual exercises, which engage students in active learning. Other research focussed on cooperative learning of statistics, which has been shown to have beneficial effects on student learning (Garfield, 1993; Giraud, 1997; Keeler & Steinhorst, 1995; Magel, 1998). Recommendations are primarily given for lectures (e.g., Larsen, 2006), methods for teaching specific subjects (e.g., Walters, Morrell & Auer, 2006), the use of data sets (e.g., Holcomb & Spalsbury, 2005), textbooks (e.g., von Hippel, 2005), and specific exercises (e.g., Kahn, 2005).

In sum, a restricted number of studies in statistics education are focussed on motivational aspects like affect and attitude. Most studies are directed at innovations in education and on the actual effects of such innovations on performance. However, modifications in education aimed at stimulating students' motivation and more specifically, aimed at creating a positive affect toward statistics, are as important as innovations directly aimed at the improvement of students' performance.

The studies of this dissertation focus not only on the effects of educational modifications on student learning, performance, and conceptual understanding, but also on some effects on motivation. The relations between education, student learning, conceptual understanding, performance, and motivation are presented in Figure 1.1. It was expected that innovations in education directly and indirectly via motivational factors, would affect student learning, which in turn was expected to affect conceptual understanding. When students would achieve conceptual understanding it was anticipated that they would perform better and probably would develop a more positive motivation toward statistics. In contrast, it was expected that when students would be unable to understand statistics, they would develop a negative motivation. Motivation was expected to affect

student learning in a circular way. The arrows in Figure 1.1 stand for these effects.

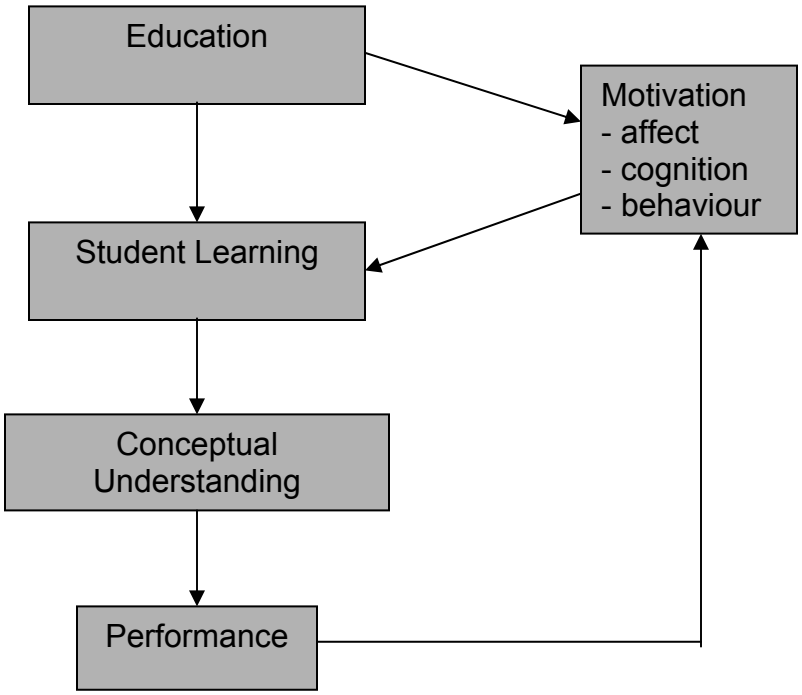



Figure 1.1. Relations between the main concepts of this dissertation.

The two educational modifications that are studied in this dissertation are directive guidance and the reduction of distributed practice. The directive guidance was investigated in three studies; two field studies and one experiment. It consisted of directive questions. These directive questions were supposed to guide the students through the subject matter and direct their reasoning processes. The reduction of distributed practice was studied in a field study, in which samples of students from two cohorts enrolled in an introductory statistics course were compared. One cohort followed the course before a curriculum change, the other after the change. The duration of the statistics course before the curriculum change was six months, after the change eight weeks. The latter cohort had considerably less opportunity to distribute study activities.

The effects of these two educational modifications on a number of motivational components, on two student learning related mechanisms, on students' conceptual understanding and on their performance were studied. The motivational components that we studied comprise 1) affective aspects, like interest and value; 2) cognitive aspects, like causal explanations and attitude; 3) behavioural aspects, like effort and persistence. The two to student learning related mechanisms that we studied are cognitive load and student reasoning.

Motivation and conceptual understanding are the two central concepts of this dissertation. Two studies are especially dedicated to these concepts. In one

study a motivational model is tested. In the other study conceptual understanding is defined and a method for its measurement is proposed and examined.

*1.1.2 Conceptual understanding*

Our focus on conceptual understanding as the outcome of student learning is in contrast to what is usual in the domain of statistics education. In statistics education the terms statistical literacy, statistical reasoning, and statistical thinking are used as possible learning outcomes. However, it is acknowledged that these concepts are unclear (Ben-Zvi & Garfield 2004; Garfield & Chance 2000). There are several reasons for this indistinctness. First, there is no formal agreement regarding the definitions of the concepts. As a consequence these definitions overlap (Broers, 2006). Chance (2002) and delMas (2002) claim that consequently the concepts are sometimes used interchangeably. In contrast to this view, Ben-Zvi and Garfield (2004) treat the concepts as hierarchically ordered.

Second, the definitions that are used are often circular and use not well defined terminology (Budé, 2006). For example, statistical reasoning is sometimes defined as: the way people reason with statistical ideas (Ben-Zvi & Garfield 2004; Garfield & Chance 2000). This is a tautology. Other definitions use terms like: the ability to understand, the ability to fully interpret and judge statistical results, and the appreciation of the concepts (Ben-Zvi & Garfield 2004; delMas, 2002; Rumsey, 2002, Wild & Pfannkuch, 1999). Such definitions do not clarify the concepts, because the terms understanding, ability to interpret, and appreciation are not strictly defined themselves. The indistinctness of the concepts statistical literacy, reasoning and thinking, leads to problems with regard to the assessment of these concepts (Broers, 2006; Budé, 2006). Hence, the focus on conceptual understanding in this dissertation. However, conceptual understanding is not a well defined concept itself. It has to be demarcated and a method for its assessment has to be determined, before it can be used as an outcome variable with which the effects of modifications in statistics education can be studied. We will give a definition and propose an assessment method in chapter 2.

*1.1.3 Overview of the dissertation*

A method for assessing conceptual understanding is described and studied in **Chapter 2**. The underpinning of this method is our definition of conceptual understanding, which is: conceptual understanding is reached when a student sufficiently knows all the relevant concepts and their mutual interrelations. From this definition follows that in our view conceptual understanding depends on the internal knowledge structures of an individual. Knowledge structures can be conceptualised as organised networks of concepts with their interrelations (Anderson, 1983; Collins & Loftus, 1975; Feltovich, Spiro & Coulson, 1993; Kintsch, 1998; Rumelhart & Norman, 1983). Rich and more interrelated knowledge structures are characteristic for conceptual understanding, superficial understanding is characterised by isolated and fragmented knowledge (Kintsch, 1998; De Jong & Ferguson-Hessler, 1996). This

means that conceptual understanding can be measured by assessing these characteristics of the knowledge structure of the individual. In order to do so, an assessment method has to measure both the relevant concepts as the relations between them. It is not sufficient to only reproduce a definition of a single concept or fill in a formula, because this can be done with isolated and fragmented knowledge. Proper assessment methods should require the correct combination of more than one concept. In this way knowledge of the concepts as well as their interrelations can be tested. This enables the assessment of the richness and the relatedness of the knowledge structures. Several methods are suitable, e.g. concept mapping, open ended questions, and writing an essay (Jonassen, Beissner & Yacci, 1993). In this dissertation two types of open ended questions are used. The first type of questions asks for relating and explaining concepts as they were presented in the course. The second type, so-called transfer questions, asks for the application of what is learned in the course, in a different example, in a different way, or in a different context. The scoring of the written answers to these open ended questions is often thought to be difficult and laborious. In chapter 2 it is shown that scoring the answers can be done reliably and validly with a detailed answer key.

In **Chapter 3** the relation between a number of motivational components and the outcome of student learning is addressed. The aspects of motivation that we studied are cognitions (thoughts), affect (feelings), and behaviour. These motivational aspects form a causal chain (Peterson et al., 1993; Pintrich & Schunk, 1996). When a person experiences a conspicuous event in a positive or negative way, she/he will attribute causes for such an event. These attributions will correspond with the positive or negative experience and they will act upon the expectations of that person. These expectations may lead to a positive or negative affect, and active or passive behaviour. This in turn may lead to a desirable or undesirable outcome.

In the context of statistics education experiencing an event can be failing or passing an exam. Attributions can be thoughts like: *I am not capable of understanding statistics*, or *statistics is easy*. Expectations in this context are for example the gain that students expect from attending a lecture or studying the literature. Affect consists of, for example, feelings of dislike or insecurity toward statistics (Gal et al., 1997). The outcome of student learning in statistics education is usually what students understand from the subject matter and their performance on the exams.

In a quasi experiment the attributions, outcome expectations, affect, and study behaviour were measured with a questionnaire during a statistics course. Performance was measured with the final exam of that course. The mediating concepts of student learning and conceptual understanding (see Figure 1.1) were not measured. A model was designed that reflected the causal chain of the motivational components. A path analysis was used to test the model. The results showed that affect plays a more central role than was anticipated. Affect seems to be the most influential motivational component with regard to the performance of students.

In **Chapter 4** the effect of directive guidance in statistics education on students' performance and a number of affective aspects toward the statistics course is investigated. In this study directive guidance was provided by the tutors in a problem based learning (PBL) statistics course. It consisted of directive questions that the tutors could use to structure the discussions in the tutorial meetings. By asking these questions, the tutors could direct the discussion in the correct way as soon as the students would go astray or omit important topics. Several studies have shown that asking questions provokes students to focus on and actively reason about the subject matter (Chi, Siler, Jeong, Yamauchi, & Hausman, 2001; Glaser, 1991; Graesser, Bagget, & Williams, 1996; Graesser, Person, & Magliano, 1995). More focussing on the subject matter and more active reasoning were expected to enhance conceptual understanding and this was expected to lead to better performance and to a more positive evaluation of the statistics course. Performance again was measured with the final course exam. A standard course evaluation questionnaire was used to measure students' appraisal of the course. The questionnaire included items regarding the motivational components value and affect. It was found that directive tutor guidance improves students' perceptions regarding value and affect toward statistics as well as their performance on the exam.

In **Chapter 5** a series of three studies is presented in which the effects of directive tutor guidance and distributed practice on students' conceptual understanding at different time points, were investigated. In the first study a change in the curriculum was examined. This change caused a reduction of duration of the statistics courses leading to more massed practice. Massed practice, relative to distributed practice, is defined as a decrease in the spacing of study activities. This massed practice was expected to have a detrimental effect on students' conceptual understanding (Bahrick & Hall, 2005; Seabrook, Brown, Solity, 2005). In the second study it was investigated whether this negative effect on students' conceptual understanding could be countered by directive tutor guidance. In contrast to the studies in chapters 3 and 4, conceptual understanding in these two studies was measured with explanation and transfer questions. In the third study of this chapter the effect of distributed practice and directive tutor guidance on long term conceptual understanding of students was investigated. Students answered the same questions as in the first two studies of this chapter, six months after the course.

The results showed that massed practice indeed leads to decreased conceptual understanding. Secondly, directive tutor guidance enhances conceptual understanding. Finally, distributed practice and tutor guidance have a positive effect on long term retention. However, the level of students' conceptual understanding six months after the course proved to be rather low.

All the aforementioned studies were done in natural educational settings. In these studies the effects of modifications of statistics education on students' conceptual understanding or their performance on the exam were investigated. It was explained that the effects of the modifications in education are mediated by student learning (see Figure 1.1). Student learning consists of complex mechanisms and processes. Asking directive questions, for example, was

supposed to focus students more on the essential topics and activate students' reasoning. Focussing students was expected to reduce cognitive load, whereas activating students' reasoning was expected to lead to more elaboration and reflection on the subject matter. However, the field studies did not allow us to investigate these mechanisms.

In **Chapter 6** an experimental study is presented in which these two to student learning related mechanisms were measured. Cognitive load is defined as the amount of effort needed by the human cognitive system to process information (Sweller, 1988). Cognitive load can be divided in extraneous load, which impedes learning and germane load that enhances student learning (van Merriënboer & Sweller, 2005). Cognitive load was measured in this study with a questionnaire. Student reasoning was measured as the successive thinking steps in self-explanations (internal explanations to oneself). In this experimental study, again students were guided by directive questions. In contrast to the former studies, written questions were used. This resulted in a more standardised form of guidance, compared to the studies in which the tutors asked guiding questions. It was investigated whether this standardised form of guidance via student reasoning and cognitive load would lead to better understanding of the subject matter and a more positive attitude toward statistics. Attitude is strongly related to affect and is defined as relatively stable automatised feelings when positive or negative emotional responses are repeated (Gal et al., 1997). As such, attitudes are considered part of a person's motivation. The written questions were supposed to enhance student reasoning and to reduce extraneous cognitive load. This in turn was expected to improve performance and lead to a more positive attitude.

The results showed that the intervention with the written questions stimulates self-explanations and improves performance without raising the cognitive load. We did also find that performance is positively correlated with attitude toward statistics.

In **Chapter 7** summaries, general conclusions, and recommendations are given. From the studies in this dissertation, it can be concluded that both directive tutor guidance as written directive questions enhance conceptual understanding and performance of the students. Second, directive tutor guidance leads to a more positive motivation. A more positive motivation in turn also leads to improved performance, which may feed back to a more positive attitude. The third conclusion is that affect is important in the improvement of statistics education. Fourth, it can be concluded that statistics education in courses of about 8 weeks limits the distribution of practice and thus conceptual understanding.

# Chapter 2

## A procedure for measuring understanding

This chapter is submitted for publication as: A procedure for measuring understanding, Luc Budé, Margaretha W.J. v.d. Wiel, Tjaart Imbos, Henk G. Schmidt, Martijn P.F. Berger

## 2.1 Abstract

*Students' understanding of subject matter is the aim of education, but difficult to define, monitor, and measure. In cognitive psychology some qualities of knowledge are attributed to understanding. Based on cognitive psychology three criteria for the measurement of understanding are proposed. These criteria are the richness of knowledge, the integration of knowledge, and the appearance of not explicitly presented knowledge. They are integrated in a procedure for the measurement of understanding. The procedure focuses on a suitable way of eliciting knowledge, representing this knowledge, and scoring. This procedure was put into practice in a study in which participants studied a science text. Results indicate that it leads to useful and reliable outcomes. Recommendations for the practice are given.*

## 2.2 Introduction

The present study is focussed on measuring understanding. Understanding of what is taught is invaluable. It may lead to better performance on inference and creative problem-solving tasks (Mayer, 1989). It may also be important for the application and transfer of knowledge (Novak, 2002). Understanding may enable knowledge to be flexibly engaged to accomplish diverse, sometimes novel objectives (Feltovich, Spiro & Coulson, 1993). Understanding may create a sensation of coherence, satisfaction, and confidence to be able to explain the subject matter (Entwistle, 1995). Furthermore, a comprehender may be able to explain causal mechanisms and causal antecedents of events and processes (Graesser, Olde & Lu, 2001; Noordman & Vonk, 1998). This is why understanding is stimulated in education and preferred over superficial learning, such as rote learning.

Measuring understanding implies its demarcation. However, it is very difficult to specify understanding. We will focus on characteristics of the knowledge that is stored in the brain of an individual. We will not discuss situated knowledge and distributed cognition (Greeno, 1998). Still it is hard to identify those features of internal, personal knowledge that enshrine understanding. This is even more complicated, because it is not known how knowledge is stored in the brain. Despite the fact that little is known about how cognition can be explained in terms of underlying brain processes and material referents of knowledge, it is possible to issue epistemological statements on how knowledge is stored in the mind (Marr, 1982). A well accepted epistemological view on knowledge concerns its representation. Knowledge is believed not be represented in an exact encyclopaedia-like way, but rather stored as elementary knowledge units in meaningful internal structures (see e.g. Anderson, 1983; Collins & Loftus, 1975; Collins & Quillian, 1969; Feltovich, et al., 1993; Glaser & Bassok, 1989; Norman, 1982; Norman, Gentner & Stevens, 1976; Quillian, 1968; Rumelhart & Norman, 1983; Schank & Abelson, 1977). We will refer to the elementary knowledge units as *knowledge elements* and their dispersion in meaningful internal structures will be called *mental structures*. The position of a knowledge element in the mental structure and the relations to other elements determine its denotation. This means that the number of the knowledge

elements, the texture, the ranking, the arrangement, and the hierarchy of the elements altogether determine knowledge. This knowledge can be correct or wrong, superficial or profound, and/or apt or inadequate in a certain context. So, the correctness, profoundness, and aptness of knowledge are dependent on the distance between knowledge elements, the location of the knowledge elements relative to each other, like a subordinate or super ordinate position, and the connections between the elements. Because such properties of the mental structures capture the meaning and the content of knowledge, they are indicative of the level of understanding. In the next section, based on these properties of the mental structures, three criteria will be proposed. These criteria should enable the differentiation between levels of understanding.

Another factor that complicates the measurement of understanding is that internal knowledge is not directly accessible. This internal knowledge has to be elicited and externally represented in such a way that the properties of the mental structures that are characteristic for understanding are preserved and revealed. In the following, suitable elicitation and representation techniques will be presented. Finally, a grading technique is needed that values those properties that are of importance.

The goal of this chapter is twofold. The main goal is to propose an operating procedure for the elicitation, external representation, and assessment of those features of internal mental structures that in our opinion characterise a deeper level of understanding. The second goal is to apply the proposed operating procedure in a study and report the findings.

### 2.3 Understanding

It is known that the mental structures of experts and novices differ (Chi, Feltovitch, & Glaser, 1981; Jonassen, Beissner & Yacci, 1993; Kraiger & Cannon-Bowers, 1995). The question is which features characterise expert knowledge; i.e. which features indicate a more profound level of understanding. We will discuss three features of internal mental structures that, according to us, point to a deeper level of understanding: the richness of the mental structures, integration of new knowledge into existing mental structures, and the presence of not explicitly presented knowledge. After these features are reified in external representations, they can be measured and used as criteria in the evaluation of understanding.

Firstly, understanding depends on the *richness* of mental structures (Anderson & Schunn, 2000; Chi, et al., 1981). Richness refers not only to the number of the knowledge elements, but in particular to their interrelations. Knowing all the concepts and their relationships enables the explanation of causal mechanisms (Wyman & Randel, 1998). Knowledge of how a system works, that is, knowing the causal relations, is characteristic for scientific understanding (Mayer, 1992). The richness of the mental structures is according to us the first distinctive feature of understanding.

Secondly, *integration* might point to a deeper level of understanding. It refers to the proximity of new knowledge elements to existing mental structures, as well as the connections between them. Kintsch (1998a) claims that it is typical

of deeper understanding when information that is presented in a text is connected to, i.e. integrated in prior knowledge forming coherent mental structures, as opposed to surface representations that consist of knowledge elements directly derived from the text. De Jong and Ferguson-Hessler (1996) also claim that deeper understanding is conditional upon new knowledge being firmly anchored in a person's knowledge base. In contrast, isolated knowledge may be typical for a more superficial sort of understanding. Rote learning, for example memorising some definition, might lead to isolated knowledge. Vosniadou and Brewer (1994) have shown that a child may have learned that the earth is a sphere, yet have a coexisting mental model of a flat earth. This proves that factual information can be learned and correctly reproduced, without being integrated in existing knowledge. Correct knowledge with coexisting misconceptions may reflect isolated, not integrated, and not understood knowledge (Novak, 2002; Wolfe, Schreiner, Rehder, Laham, Foltz, Kintsch & Landauer, 1998). Therefore, only when correct knowledge is shown to be integrated in existing knowledge, proof of true understanding is produced. Thus, integration can be seen as a second characteristic feature of understanding.

A third characteristic of understanding might be the appearance of *not explicitly presented* knowledge. Glaser (1993), like Kintsch (1998a), makes a distinction between a surface representation and a deeper level of representation. The surface representation consists of knowledge that was explicitly presented during learning. Glaser asserts that the deeper level of representation is characterised by an organisation of knowledge that includes additional implicit concepts, with their relationships, and the principles pertaining to these relationships. Whenever information is passed on during learning there are always some matters left implicit. If it can be shown that the learner has filled in these missing knowledge elements in his mental structure, this might be symptomatic for a deeper level of representation. According to Glaser, this deeper level of representation is typical for experts and successful problem solvers. Additional, not explicitly presented knowledge, therefore, is regarded as the third characteristic of a more profound level of understanding.

To be able to evaluate these properties of mental structures, so as to determine the level of understanding, calls for an appropriate way of eliciting and representing these mental structures.

### 2.4 Elicitation and external representations of internal mental structures

Three criteria need to be evaluated for the determination of the level of understanding. These criteria are: the richness of the mental structures, the integration of new knowledge elements in existing mental structures, and not explicitly presented knowledge elements in the mental structures of an individual. The three criteria are not tangible; they are properties of internal knowledge, which is not directly accessible. To evaluate these three properties of the internal mental structures, they somehow need to be preserved during elicitation and external representation. This implies that elicitation of knowledge should be aimed at the relations of knowledge elements and not at recognition and word-for-word recall of information as it was presented (Feltovich, et al., 1993).

Recognition and word-for-word recall of information will only reveal isolated knowledge. Elicitation should explicitly be aimed at the combination of several elements of the domain, at the underlying causal mechanisms, and at the connection of conditions and procedures for their application (Dochy, 2001; Gijbels, Dochy, van den Bossche & Segers, 2005). In this way the relationships between the knowledge elements can be revealed and by that the richness of the knowledge. There are several appropriate kinds of elicitation techniques: verbal and written tests (Jonassen, et al., 1993), word association, proximity ratings (Schvaneveldt, 1990), similarity ratings (Goldsmith, Johnson & Acton, 1991), card sorting, and concept mapping (Ruiz-Primo, Schultz & Shavelson, 2001). Standardised tests like multiple choice or true/false tests may not be suited for eliciting all properties of knowledge (Dochy, 2001). For example, they might not adequately reproduce the integration of knowledge, or the presence of additional knowledge. After being appropriately elicited, the knowledge has to be represented in an adequate way.

External representations depict knowledge as long as they reflect the relevant features of the elicited internal mental structures (Kintsch, 1998b; Olson & Biolsi, 1991). Suited modes of external representation are: written answers (Jonassen, et al., 1993), networks of propositions, networks of concepts, tree representations, semantic vectors (Landauer & Dumais, 1997), concept maps, multi-dimensional scaling (Kruskal, 1964), and pathfinder networks (Schvaneveldt, 1990).

In the following, we will discuss elicitation, representation, and grading techniques that can appropriately be used in a procedure to measure understanding. More specifically, explanation and transfer questions will be presented as suitable techniques for the elicitation of knowledge. Network representations and written answers will be presented as suitable techniques for the representation of knowledge.

In educational settings explanation questions are regularly used for eliciting knowledge. To answer explanation questions, one has to explain relations and causal mechanisms of concepts that are studied. It is not sufficient to give definitions or verbatim quote presented information, demonstrating isolated knowledge. Concepts need to be associated with each other and it has to be explained how they are related to each other. Therefore, they are suited for revealing the richness of the mental structure of the learner. Moreover, a learner has to explain the topics in her/his own words, and by that displaying the integration of her/his knowledge. This is in contrast to, for example, answering some multiple choice questions. Many multiple choice questions can be answered with isolated knowledge. To use your own words and to be able to provide self contrived examples and descriptions is only possible if integration has taken place. So, both rich knowledge structures as well as integrated knowledge are needed to be able to answer explanation questions (Mannes & Kintsch, 1987). Transfer questions, similar to transfer problems, are a particular type of explanation questions. They are not so much concerned with an accurate explanation of presented information, but more directed at applying the knowledge to a situation not explicitly explained in the studied material.

Answering transfer questions asks for knowledge about procedures, principles and conditions of the correct application of the learned subject matter. This application can relate to the same procedures / principles / conditions in different combinations (near transfer), the same procedures / principles / conditions in a different context (far transfer), or novel procedures / principles / conditions in a novel context (very far transfer) (Bassok, 2003; Campione & Brown, 1990; Phye, 1992; Reimann & Schult, 1996; Renkl, Mandl & Gruber, 1996). This information about procedures, principles and conditions is usually not explicitly presented. Transfer questions, therefore, are suited for the elicitation of not explicitly presented knowledge. As such they are especially suited to distinguish students of varying academic ability (Bassok, 1997; Mayer, 1997; Mayer & Wittrock, 1996). Although explanation questions and transfer questions are focussed at slightly different aspects of knowledge, they both are regarded suitable elicitation techniques in a procedure to measure understanding.

The written answers to such questions are already appropriate external representations of internal knowledge (Jonassen, et al., 1993; Olson & Biolsi, 1991). However, it is problematic to obtain objective scores from such answers (Kintsch, 1998a). Moreover, as explained, the evaluation should be focussed at the richness of the knowledge (number of knowledge elements and relations), at the integration of knowledge (connection of new knowledge with prior knowledge), and at not explicitly presented knowledge. Network representations visualise these features of knowledge and enable a precise, extensive and objective scoring. As such these network representations preserve and enable the objective scoring of those features that are characteristic for deeper levels of understanding. They are, therefore, suited representation techniques in a method for the measurement of understanding. In our operating procedure we propose two of such representation techniques: Pathfinder networks and Semantic networks.

The Pathfinder software generates network representations using proximity ratings. In these networks (Pf-nets) items are represented as nodes and relations as links (Cooke, 1992; Dearholt & Schvaneveldt, 1990; Goldsmith, et al., 1991). Pf-nets have shown to be informative about memory organisation (Cooke, Durso & Schvanefeldt, 1986), can be used to infer knowledge patterns from action patterns (Rowe, Cooke, Hall & Halgren, 1996), and may yield a valid structural assessment of classroom learning (Goldsmith & Johnson, 1990). In addition, Pathfinder proximities may represent psychological proximities (Cooke, 1992). Semantic networks (Semnets) are handmade network representations. These networks also consist of nodes and the links between them. The nodes represent concepts and the links represent the relations between these concepts. Semnets can be used to represent knowledge, and differentiate between different levels of expertise (Schmidt & Boshuizen, 1993; Van de Wiel, Boshuizen & Schmidt, 2000).

### 2.5 Research questions

In this study an operating procedure for measuring what participants understood of a science text was put into practice. The procedure focussed on

the three proposed criteria, 1) by using explanation and transfer questions that elicited internal knowledge in an appropriate way, 2) by using written answers, Pf-nets and Semnets as suitable external representations, 3) and by using four complementary scoring techniques.

- The primary research question was whether this procedure would yield a reliable measurement of understanding.
- A second question was how the answers on the explanation questions would relate to the answers on the transfer questions. To answer this question a factor analysis was performed on the separated scores for the answers on the explanation and transfer questions. Two positively correlated factors were hypothesised, because the two types of questions focus at different aspects of knowledge, yet both ask for deeper understanding.
- Our final question was whether we would find a low positive correlation between participants' estimations of their own understanding and performance on a test. The correlation between subjective measures and other measures of understanding is known to be low (Glenberg, Sanocki, Epstein & Morris, 1987).

## 2.6 Method

### *2.6.1 Participants*

Twenty first-year psychology students, randomly selected from a volunteer pool of students, participated in this experiment. Two were male and eighteen female. Nine attended physics courses in secondary school. All participants received course credits. To motivate the participants to answer the questions to the best of their capabilities, an extra €5 reward was offered to those five participants who performed best.

### *2.6.2 Materials*

An instructional text of four pages about the science class topic of thunder and lightning was used. The text was comprehensible with only little prior knowledge. The development of thunder and lightning was explained, as well as the mechanisms that lead to discharges and the risks involved with lightning. The topic enabled questions about the relations between the phenomena, explanations of the phenomena, and causal mechanisms.

Ten questions were formulated about the phenomena that were discussed in the text. The first six questions were explanation questions. In order to answer these questions correctly, participants had to understand and explain the relations between concepts. Four questions were transfer questions. The information required to answer these transfer questions, was not mentioned explicitly in the text. One was related to lightning (far transfer); three were about general electric phenomena (very far transfer).

For the scoring of the answers four tools were developed.

1. Standard answers, based on the text, were constructed and appraised by two experts, secondary school physics teachers. These standard answers are written answers that consist of comprehensive explanations of the phenomena. This tool

was used for constructing an answer key, for model semantic networks, and for appointing proximity ratings, needed for the model Pf-net.
2. An answer key was developed for each question by ascribing one point to each of the main concepts in the standard answers.
3. Semnets were constructed, representing the main concepts and the relations between these concepts. For each question a network was constructed to serve as a model.
4. Model Pf-nets were constructed for each question using the pathfinder software and proximity ratings (see paragraph 2.6.4).

A questionnaire was developed to assess participants' estimations of their own understanding. The questionnaire consisted of ten statements, to which participants could respond to on a seven-point Likert scale (1 = totally disagree, 7 = totally agree). Six questions were concerned with the participants' estimation of their performance with regard to the explanation and transfer questions. Four questions were concerned with the participants' estimation of their general comprehension of the text.

*2.6.3 Procedure*

Participants were given written instructions about their task. In addition it was orally emphasised to study the text carefully and answer the questions in complete sentences. Participants were encouraged to make notes in the text and on a blank sheet. Next, all of them studied the text for 40 minutes. After returning the text and their notes to the experimenter, 60 minutes were available for answering the questions. Participants used 30 up to 50 minutes to complete this task. Finally, they filled in the questionnaire, using about 10 minutes.

*2.6.4 Scoring methods*

Four scoring methods were used to assess participants' answers: 1) scores derived from a global intuitive evaluation of two physics teachers, 2) scores derived from two other raters using a detailed answer key, 3) scores derived from the same two raters using the Semnets, 4) scores derived from Pf-nets.
1. The teachers' scores focussed on global aspects. The score for each of the ten answers varied from ten points when a subject showed good understanding of the essence, to zero points when there was no understanding at all. An overall grade was calculated by adding up the question scores divided by ten.
2. A second score was determined by using the answer key. This tool was developed by ascribing one point to each of the main concepts in the standard answers. The score per question could vary from zero to twenty points. Every main concept of the answer key that was mentioned in the answers of the participants was counted and awarded with the corresponding point. In contrast to the global intuitive evaluation of the first method, the purpose of this method was to obtain a more precise and objective score. In this measure there was no concern for the relations among the concepts, only the number of main concepts were included. For an overall grade, question scores were added up.

After completing this scoring procedure an additional rater scored the answers using the answer key, for interrater reliability purposes. This rater was a complete outsider, who had not read the text or seen the questions.

3. The third scoring method used the Semnets. First the concepts and links that the participants mentioned in their answers were mapped onto the Semnets which served as models (Van de Wiel, et al., 2000). Next the concepts and links were counted and evaluated. Nine aspects were scored:

A. The total number of mentioned concepts.
B. The total number of mentioned links.
C. The number of concepts matching the model concepts.
D. The number of links matching the model links.
E. The number of extra correct concepts.
F. The number of extra correct links.
G. The number of shortcuts due to concise but correct answers.
H. The number of all possible model concepts.
I. The number of all possible model links.

The accuracy of the answers was determined by two ratios. The first ratio was: $R_1 = (C + E) / A$. The second ratio was: $R_2 = (D + F + G) / B$. The completeness of the answers was determined by two other ratios. These ratios were: $R_3 = (C + E) / H$ and $R_4 = (D + F + G) / I$. In establishing the quality of the participants' answers the four ratios are complementary. Taken together they express how much of the answers is correct and how complete the answers are. An overall grade was calculated by summing up the ratios of all questions.

4. Scores derived from Pf-nets focussed on the proximity of the concepts in the written answers. Pathfinder generates graph structures based on proximity data. To construct Pf-nets, the answers had to be converted into proximity data. For this purpose the following procedure was followed. First the main concepts in the standard answers were determined. Next pair-wise distances between concepts in the participants' answers were established. This was done by assigning cumulative points if concepts were: both mentioned in the answer (2 points), in consecutive sentences (3 points), in the same sentence with other concepts in between (4 points), in the same sentence with one concept in between (5 points), adjacent concepts (6 points). If concepts were mentioned more than once additional points were only awarded if concepts were more than once in the same sentence. The total of the points was cut off at a maximum of nine. A high score thus reflects a high relatedness. The obtained distances were entered into a rating program as proximities. These proximities were subsequently used to construct Pf-nets. The *r* parameter that is used in the computation of the length of paths was set on infinite. The *q* parameter that specifies a limit on the number of links was set on *n*-1 (*n* is the number of nodes, which stands for the number of concepts). The standard answers were used to generate model Pf-nets. The Pf-nets generated from the participants' answers were compared to the model Pf-nets. The similarity between model Pf-nets and participants' networks was assessed by considering both concepts as well as links (Schvaneveldt, 1990).

The similarity could vary from zero to one for each question. An overall grade was calculated by summing up the similarities of all questions.

Combined these four scoring techniques reflect all the relevant properties of the participants' mental structures, i.e. the number of the knowledge elements (concepts), the content of the concepts, the number of the relationships, the content of the relationships, the correctness of the relationships and the proximity of the concepts. Therefore, these four techniques together are considered to constitute an elaborate and complete assessment of participants' knowledge. For this reason they were summed up to obtain an overall grade, which we assume to be a valid measure of what the participants understood of the text.

*2.6.5 Analysis*

First the results of the four scoring techniques were converted into four variables in the following way. The first scoring technique consisted of two physics teachers who graded the answers. For the interrater reliability the means and a correlation was computed. The means show that one of the teachers gave lower grades with a smaller standard deviation ($M_1$ = 5.02, $SD_1$ = 1.97; $M_2$ = 3.92, $SD_2$ = 1.46). However, this was done consistently ($r$ = .914, $p$ < .01). Their results were averaged for further analysis into one variable called *Teach*. The second scoring technique consisted of two of the raters using the answer key. Both correlation ($r$ = .985, $p$ < .01) and the means and standard deviations ($M_3$ = 3.37, $SD_3$ = 1.33; $M_4$ = 3.00, $SD_4$ = 1.22) showed a high interrater reliability for these raters. Their results were also averaged for further analysis into one variable *Key*. The interrater reliability of the raters using the Semnets was also high ($r$ = 0.896, $p$ < .01); the means and standard deviations of the grades were similar ($M_5$ = 4.91, $SD_5$ = 1.76; $M_6$ = 5.05, $SD_6$ = 1.84). Their results were also averaged. This variable is called *Semnet*. The grades that resulted from the scoring method using the Pf-nets were called *Pfnet*.

These four variables (*Teach, Key, Semnet,* and *Pfnet*) were added up to obtain an overall grade. The scores of the overall grade constitute the variable *Overall*. As was expected, the subjective rating (*Subj*) correlated low with the other four variables. For this reason it was not used in the construction of the overall grade. For the six variables *Teach*, *Key*, *Semnet*, *Pfnet*, *Subj*, and *Overall* correlations were computed.

For a factor analysis the scores of the four scoring techniques were separated for the two types of questions, i.e. explanation and transfer questions. This yielded eight variables: *Teachexpl*, *Keyexpl*, *Semnetexpl*, *Pfnetexpl*, *Teachtrsf*, *Keytrsf*, *Semnettrsf*, and *Pfnettrsf*. For these variables correlations were computed and a Principal Axis Factoring (PAF) was done to obtain factors that optimally explained the interrelations between the variables. PAF uses the communalities of the variables and is therefore more suited for this purpose. An oblique rotation technique (Oblimin) was chosen, because it was expected that the factors would be correlated. This yielded clearer patterns of factor loadings and better interpretable factors.

## 2.7 Results

Table 2.1 presents the correlations between *Teach*, *Key*, *Semnet*, *Pfnet*, *Subj*, and *Overall*, as well as the means and standard deviations of these variables. All correlations were significant ($p < .01$).

Cronbach's α was computed for the four variables that were combined for the overall grade ($\alpha = .9772$). The interrater reliability for the scoring with the answer key for two independent researchers compared to a lay rater was high ($r = .982$ and .978).

Besides computing the correlations between each of these four variables and the variable *Overall* (which is composed of these four variables), correlations were also computed for each variable separately with a combination of the other three variables. These correlations are presented in Table 2.1 as scoring method rest correlations.

Table 2.1. Correlations between the six variables, their means and standard deviations, and scoring method rest correlations.

| | *Teach* | *Key* | *Semnet* | *Pfnet* | *Subj* | *Overall* | *Scoring method rest correlation* [a] |
|---|---|---|---|---|---|---|---|
| *Teach* | 1.00 | .944 | .855 | .857 | .565 | .976 | .943 |
| *Key* | | 1.00 | .890 | .899 | .587 | .983 | .955 |
| *Semnet* | | | 1.00 | .850 | .629 | .921 | .883 |
| *Pfnet* | | | | 1.00 | .442 | .926 | .892 |
| *Subj* | | | | | 1.00 | .584 | |
| *Overall* | | | | | | 1.00 | |
| *Mean* | 4.47 | 3.18 | 4.98 | 3.75 | 5.12 | 4.10 | |
| *S.D.* | 1.68 | 1.27 | 1.29 | 1.32 | 1.51 | 1.39 | |

Note: [a] Scoring method rest correlations were computed between respectively: *Teach* and (*Overall* minus *Teach*), *Key* and (*Overall* minus *Key*), *Semnet* and (*Overall* minus *Semnet*), and *PFnet* and (*Overall* minus *Pfnet*).

Table 2.2 presents the correlations between the variables *Teachexpl*, *Keyexpl*, *Semnetexpl*, *Pfnetexpl*, *Teachtrsf*, *Keytrsf*, *Semnettrsf*, and *Pfnettrsf*.

A factor analysis was done. Two factors had Eigenvalues over 1.00. The scree plot also pointed to a two factor solution, the third factor had an initial Eigenvalue of 0.563. All variables related to explanation questions loaded on one factor (*Explanation*), all other variables loaded on the other factor (*Transfer*). The results of the factor analysis after rotation are presented in Table 2.3.

Table 2.2. *Correlations between the separated scores for explanation and transfer questions.*

| | *Teachexpl* | *Keyexpl* | *Semnetexpl* | *Pfnetexpl* | *Teachtrsf* | *Keytrsf* | *Semnettrsf* | *Pfnettrsf* |
|---|---|---|---|---|---|---|---|---|
| *Teachexpl* | 1.00 | .911** | .896** | .818** | .477* | .530* | .370 | .404 |
| *Keyexpl* | | 1.00 | .901** | .889** | .493* | .500* | .436 | .381 |
| *Semnetexpl* | | | 1.00 | .830** | .465* | .508* | .499* | .401 |
| *Pfnetexpl* | | | | 1.00 | .298 | .341 | .322 | .206 |
| *Teachtrsf* | | | | | 1.00 | .891* | .751** | .773** |
| *Keytrsf* | | | | | | 1.00 | .786** | .721** |
| *Semnettrsf* | | | | | | | 1.00 | .696** |
| *Pfnettrsf* | | | | | | | | 1.00 |

Note: * $p < .05$. ** $p < .01$.

Table 2.3. *Factor analysis on separated scores for explanation and transfer questions, and the subjective ratings.*

| | Factor | |
|---|---|---|
| | *Transfer* | *Explanation* |
| *Teachexpl* | .081 | .891 |
| *Keyexpl* | .079 | .941 |
| *Semnetexpl* | .116 | .878 |
| *Pfnetexpl* | -.152 | .974 |
| *Teachtrsf* | .941 | -.023 |
| *Keytrsf* | .874 | .047 |
| *Semnettrsf* | .861 | -.006 |
| *Pfnettrsf* | .854 | -.069 |
| *Subj* | .624 | .116 |
| | | |
| *Eigenvalues* | 5.397 | 1.710 |
| *% of Variance Explained* | 59.96 % | 19.00 % |

Note: Extraction method: Principal Axis Factoring, Rotation method: Oblimin with Kaiser Normalisation

There was a positive correlation between the factors ($r$ = .491) after rotation.

## 2.8 Discussion and conclusion

The measurement of what students have understood from subject matter is far from straightforward. First, it is necessary to define and demarcate the fuzzy concept of understanding. Secondly, it is difficult because the level of understanding is a quality of internal knowledge. Even when aspects of that internal knowledge have been defined as characteristic for a profound level of understanding, some appropriate way of elicitation and representation has to be established. Finally, it is troublesome because an appropriate way of scoring has to be chosen. Yet, a more profound level of understanding is strived for in most educational settings. These facts illustrate the need for a reliable and valid method to measure what learners have understood from material that they studied.

In this paper three features of internal knowledge are proposed as criteria for the determination of the level of understanding: the richness of mental structures, the integration of new knowledge into existing mental structures, and the presence of knowledge that was not explicitly presented. An operating procedure to measure understanding by focussing on these criteria consisted of explanation and transfer questions as elicitation technique, written answers,

Semnets, and Pf-nets as external representations, and four complementary scoring methods.

The explanation questions in this study elicited participants' knowledge in such a way that the richness and integration could be evaluated. These questions obliged the participants to explain relations and causal mechanisms. This revealed the richness of their knowledge. It was not possible to literally quote passages of the text, because the restricted time to study the text did not enable the participants to verbatim memorise complete sections of that text. As a consequence, participants had to explain the phenomena using their own words. This demonstrated the integration of the new knowledge. The transfer questions asked for explanations, which were not presented in the studied material. This revealed knowledge that was not explicitly presented. Together, the explanation and the transfer questions revealed what the participants had understood. They reflect the richness of the knowledge, integration, and not explicitly presented knowledge, i.e. they elicited the relevant aspects of their mental structures.

Moreover, factor analysis showed two clear-cut factors. The scores of explanation questions loaded on one factor, the scores of the transfer question onto the other. This suggests that there is a separate aspect of understanding that is tapped by transfer questions. This aspect might be typical for the deepest level of understanding (Bassok, 1997; Campione & Brown, 1990; Mayer, 1997; Mayer & Wittrock, 1996). Our results may demonstrate the usefulness of transfer questions, i.e. open ended questions where learners have to apply what they have learned in a way different from how it was presented.

Written answers to open ended questions are a form of external representations. Written answers retain the structural properties of students' internal mental structures, provided that students are asked to relate concepts, to describe the relationships and explain the nature of these relationships. They can directly be used for assessing structural knowledge by scoring, for example, the number of correct classifications, rank judgments, etc (Jonassen, et al., 1993). Yet, in order to more clearly reveal the number of correct and incorrect concepts, number and content of the relations between those concepts, and the relatedness (proximity) of the concepts, the answers were mapped onto model Semnets and were converted into Pf-nets. These procedures resulted in three kinds of external representations: written answers, Semnets, and Pf-nets, that preserved those aspects of internal knowledge that are characteristic for understanding. These external representations had to be scored subsequently.

Four scoring techniques were used in this study. A classical scoring method (*Teach*) focussed on a global intuitive impression of the answers. The scoring method using the answer key (*Key*) focussed solely on the number of knowledge elements. The Semnets scoring method (*Semnet*) focussed on the number and content of both knowledge elements as their relations. The Pf-nets scoring method (*Pfnet*) was innovative. The distances between the knowledge elements in the participants' answers were computed and converted into proximity matrices which in turn were used to produce Pf-nets. The focus in this scoring method was thus on the proximity, reflecting the relatedness, of the

knowledge elements in the answers. Despite these differences between the four scoring methods they strongly correlated (see Table 2.1).

The focus of the four scoring methods on different aspects of the structural aspects of the answers makes them complementary. Therefore, they were combined into an overall classification of the level of understanding, the overall score. The overall score reflects most of the relevant aspects of the participants' knowledge structures, and is consequently the best reflection of their understanding of the studied material.

When the four scoring methods are considered to be four parts of one test, their intercorrelations can be interpreted as a measure of the reliability. For this reason Cronbach's alpha was computed over these four methods. The result ($\alpha$ = .9772) shows that the combination of the four methods yields a reliable measure.

Alternatively, it can also be claimed that the four methods are each individually suited for scoring written answers in order to measure understanding. Based on the three from cognitive psychology derived criteria for understanding, based on what specific aspects of knowledge are elicited by explanation and transfer questions, and based on what is being assessed by the four scoring techniques, we assume that the overall score (*Overall*) is a valid portrayal of what participants understood of the text. Given that all four separate scoring methods are highly correlated with this overall score, each scoring method could be seen as an autonomous assessment technique. If the four scoring techniques are considered concurrent assessments, than the correlations between the four scoring methods can tentatively be interpreted as an indication of the validity of each scoring method. Moreover, all methods are highly correlated with a combination of the other three methods. Taken together, all four methods seem to be just as valid as the overall score, because of the high intercorrelations. However, to truly establish the validity of any of these methods, these results should be corroborated by validation using external data. Tentative interpretations toward validity are in this study only possible on account of the above-mentioned assumption that the overall score measures levels of understanding.

It should be noted that the scoring with an answer key is the least laborious technique and it has the highest correlation with both the overall score and the other techniques. Moreover, the interrater reliability between the scoring of a lay rater, who had not read the text, nor seen the answers, and two researchers using the answer key, was high. Hence, scoring with an answer key can reliably be done by a lay rater. Using the answer key might be the most practical method for scoring written answers, because it is the least laborious method, and because a lay rater can use it. It seems to be a reliable method because of the high correlations with the other scoring methods, and because of the high interrater reliabilities. It might also be a valid way to score the answers because of the high correlation with the overall score.

The operating procedure as it was proposed in the introduction is based on what is known about how knowledge is stored in the mind. In this study this procedure was put into practice. Results show that it seems to be a useful

approach in the search for a practical, reliable, and valid method to measure what learners understand from presented material.

A final finding in this study confirms previous research. The correlation between subjective assessment of knowledge and performance on a test of that knowledge is known to be low (Glenberg, et al., 1987). In this study this was confirmed, the subjective rating correlated markedly less with the other variables than the intercorrelations of all other variables.

Future research is needed to confirm the findings in this study. First, this was a rather small study. Second, other ways of validation are needed, because in this study the correlations between four scoring methods of the same answers were interpreted as an indication of the validity. The results of a measurement using the answer key could for example be compared with other achievements in the same domain. Finally, this study was done in the domain of science. In this domain phenomena are related. Cause and effect relations in this domain make it suited for the procedure to measure understanding as it was followed in our study. In future research it could be studied whether this approach is applicable in other domains and other settings as well.

# Chapter 3

## Students' achievements in a statistics course in relation to motivational aspects and study behaviour

This chapter is published as: Students' Achievements in a statistics course in relation to motivational aspects and study behaviour, Luc Budé, Margaretha W.J. v.d. Wiel, Tjaart Imbos, Math. J.J.M. Candel, Nick, J. Broers, Martijn P.F. Berger (2007). Statistics Education Research Journal, 6 (1).

## 3.1 Abstract

*The present study focuses on motivational constructs and their effect on students' academic achievement within an existing statistics course. It is claimed that motivation is an important factor in statistics courses. Therefore, it is important to study motivation when engaged in reforming statistics education. First year Health Sciences students, participating in an introductory statistics course, filled in a questionnaire that measures several motivational constructs, namely: dimensions of causal attributions, outcome expectancy, affect, and study behaviour, all with respect to statistics. The results showed that when the cause of negative events was perceived as uncontrollable, outcome expectancy was negative. When the cause of negative events was perceived as stable, affect toward statistics was negative. Furthermore, the results showed that negative affect toward statistics and limited study behaviour, led to unsatisfactory achievements. Path analysis (Lisrel) largely confirmed the causal relations in a model that was based on attributional and learned helplessness theories. The consequences of these findings for statistics education are discussed.*

## 3.2 Introduction

Motivation influences the scope and the quality of study behaviour of students (see e.g., Bruning, Schraw & Ronning, 1999; Deci & Ryan, 1985; Graham & Weiner 1987; Pintrich, 2000). High-quality study behaviour involves active knowledge construction. Active knowledge construction is known to enhance understanding of the material in many courses (see e.g. Chi, de Leeuw, Chiu, & LaVanger, 1994; Phye, 1997; Steffe & Gale, 1995), including statistics courses (see e.g., Garfield, 1993; Giraud, 1997; Keeler & Steinhorst, 1995; Magel, 1998). Therefore, in attempts to improve statistics education, it is inevitable to stimulate motivation.

Research on motivation is quite extensive and it covers heterogeneous constructs (see e.g., Ames, 1992; Boekaerts, 1997; Volet, 1997; Weiner 1992). Some of these constructs involve phenomena that are difficult to change, because they are to a large extent determined by traits of the individual that is involved, e.g., goal orientation, self-determination, and competence. Our aim is not to focus on such phenomena, but rather to focus on constructs that have practical implications for statistics education, i.e. constructs that can be manipulated and acted upon while trying to improve statistics education.

For that reason we have focussed on two motivational theories that offer opportunities to intervene in motivational processes. Both theories take as starting-point the explanations people perceive for events they experience. These so called causal explanations have cognitive, affective and behavioural consequences. Examples for a statistics educational context of cognitive consequences are expected outcomes of visiting lectures or studying a course book, examples of affective consequences are enjoyment, pleasure, and interest, and examples of behavioural consequences are effort and persistence. The influence of causal explanations on cognition, affect and behaviour might be manipulated and driven toward outcomes that are more positive, in terms of motivation. As a consequence, these causal explanations have practical

implications for statistics education, because the obtained improvement of motivation might result in study behaviour that enhances understanding. The goal of the study was to investigate these phenomena in the context of statistics education.

### 3.3 Motivational model

In statistics education one can sometimes encounter students who think that there is a stable cause for failing an exam (e.g. statistics is a difficult subject). These students may no longer expect to benefit from studying statistics; they may start to dislike it and will not spend much study time on this subject. Other students may think that they have no control over the outcomes of their actions. For example, “no matter how hard I study, I will not be able to understand it”. These students may in advance expect to fail on the exam, will also start to dislike statistics and will not spend much time studying it. These examples show the influence of causal attributions (stability of causes, non-controllability of causes) on cognitions such as outcome expectancies (not to benefit from studying statistics, expectancy to fail on the exam) and consequently on emotions (affective reactions of starting to dislike statistics) and behaviour (disregarding statistics), which will finally have an effect on achievement. This chain effect, which is consequential for statistics education, is reflected in a model that was developed and tested in this study.

The model as a whole stands for motivation (see Figure 3.1). Motivation is not a separate entity in our model for two reasons. Firstly, it is difficult to insert it separately into a model, because it is an abstract, complex (Weiner, 1992), and ill-defined (Murphy & Alexander, 2000) construct, which is frequently used in colloquial language and consequently has several connotations. Moreover, motivation is studied in different domains and from different perspectives, which has led to distinct and changing conceptualisations and approaches. Various motivational constructs are studied, e.g. self-efficacy, goal orientation, metacognitive strategies, value, strategy use, causal perceptions, autonomy, social relatedness, etc. (see e.g., Ames, 1992; Boekaerts, 1997; Dweck, 2000; Pintrich & Schunk, 1996; Volet, 1997; Weiner, 1986). In these studies it is often left implicit whether these constructs are part of motivation or are merely related to motivation (Murphy & Alexander, 2000). Our model as a whole reflects our perspective on motivation.

Secondly, it is in our view not necessary to integrate motivation as a separate construct in the model. Traditionally, motivation was seen as an isolated latent construct that drives behaviour, cognition, and affect. We think that motivation merely is the sum of behaviour, cognition, and affect. Our opinion is in accordance with the remark of Weiner (1992), referring to Kelly (1958), that motivation as a model construct might be redundant; it is sufficient to represent only those variables that make up motivation. This view is also compatible with the fact that most motivational models do not explicitly contain motivation as a construct (see e.g., Bruning, et al., 1999; Deci & Ryan, 1985; Pintrich, 2000; Pintrich & Schunk, 1996; Weiner, 1992). Therefore, the model that we developed

contains only manifest variables, which altogether stand for motivation and it does not contain motivation as a separate latent entity (see Figure 3.1).

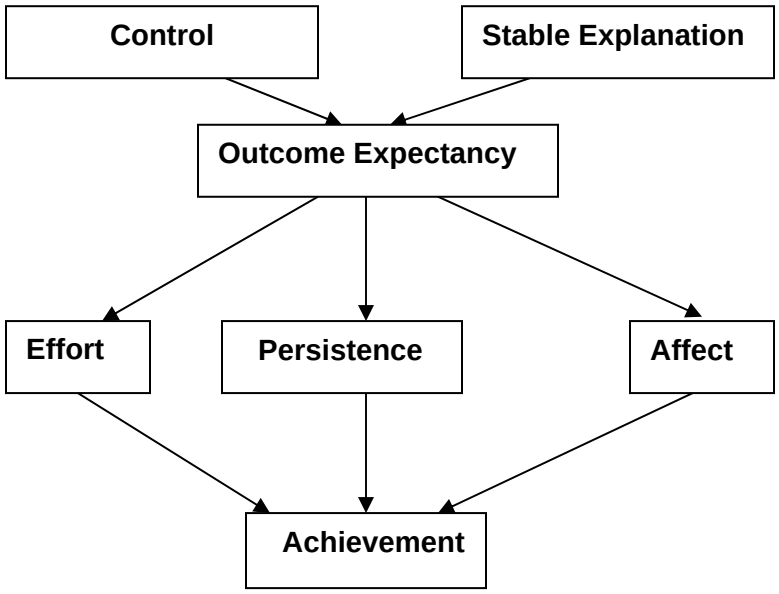



Figure 3.1. Statistics motivational model based on the attributional and the learned helplessness theory.

Two specific motivational theories were used for this model; namely the attributional and the learned helplessness theory, because they both take as starting-point perceived causes for aversive events. The attribution-based theory of motivation (Graham & Weiner, 1987; Pintrich & Schunk, 1996; Weiner, 1986, 1992) commences with perceived causes for failure, unexpected outcomes, unusual events, and important situations. Perceived causes are the way people explain to themselves such outcomes, events, and situations. The connotations of the explanations are determined by underlying properties. In attribution-based theory these underlying properties of such explanations are divided into three dimensions: stability, control, and locus. Pintrich and Schunk (1996) propose, however, that the stability dimension is most closely linked to beliefs regarding future success (outcome expectancy) and subsequently to affect and actual achievement behaviour. Therefore, we integrated *stable explanation* in our model in Figure 3.1. It can be defined as the invariability over time of such perceived causes, i.e. causal explanations.

Peterson, Maier, and Seligman (1993) present a motivational theory, which originally emanates from the learned helplessness paradigm. In this paradigm, individuals are thought to become passive and to develop affective deficits if they cannot control and avoid the causes of aversive stimuli. They claim therefore, in contrast to Pintrich and Schunk (1996), that controllability is the major factor contributing to a negative outcome expectancy. Uncontrollable

events will, according to Peterson et al. lead to a perceived non-contingency between persons' actions and the outcomes of their actions. This negative outcome expectancy will lead to pessimistic thoughts, negative emotions (affect), and passivity (behaviour). This is what is called learned helplessness. We integrated *control* influencing outcome expectation as a separate construct in our model (see Figure 3.1). Control is defined as the ability to avoid the causes of aversive stimuli.

Although the two presented theories slightly differ in the emphasis of the causal dimensions control and stability, they both reflect the way how these properties of negative causal explanations contribute to a negative *outcome expectancy*, how this will act upon *affect* and on behaviour, such as *effort* and *persistence,* which will finally result in an effect on *achievement*. The causal relations among these constructs are symbolised by arrows in our model that is presented in Figure 3.1.

This model was examined within the domain of statistics education. This means that all the constructs were measured with respect to statistical events and phenomena. It is known that perceived causal explanations via expectancy, affect, and behaviour determine future achievements in mathematics (see e.g. Seegers & Boekaerts, 1993; Vålas & Søvik, 1993). Our question was whether this is also true for statistics education and if the results would provide useful information for the improvement of statistics education. The research questions that were addressed are:

1. How do students causally explain statistics related events. Do they think that they have control over, for example, the mastery of the material, the amount of time they can spend on studying statistics, and the result on the tests? We also wanted to know whether the causes that the students reported for these events were stable or not.
2. We further measured the outcome expectancies, i.e. whether students experience a contingency between studying statistics and their understanding of the topics and the grades they receive at statistics tests. We also investigated the influence of outcome expectancy on effort, persistence and affect.
3. Finally, we investigated the relations between these motivational constructs and achievement. The potential causal relations among these constructs were tested with structural equation modelling via Lisrel (Jöreskog & Sörbom, 1989).

## 3.4 Method

### *3.4.1 Participants*

Two hundred (*N* = 200) first year students of the faculty of Health Sciences participated in a pilot study to establish the reliability of a questionnaire that was developed to measure the motivational constructs. In the subsequent year ninety-four (*N* = 94) first year students of the faculty of Health Sciences participated in the main study; 79 of these participants were female, 15 were male. The age ranged from 19 to 26 years. Approximately 75 percent of the first year Health Sciences students is female. The participants were recruited during educational activities before the start of the introductory statistics course in which this study was executed. During recruitment they were told that they had to fill in

questions about statistics education and that they would be paid 10 euro. This payment was given to avoid attracting only motivated students who were particularly interested in statistics. All participants took the introductory statistics course.

*3.4.2 Measurement procedure*

A questionnaire to measure the motivational constructs that are relevant for our model was developed. This Motivation toward Statistics Questionnaire (MSQ) consisted of 38 items, divided into six subscales. The items were phrased as statements and participants responded on a 7-point Likert scale. The questionnaire is partly a Dutch translation of the Survey of Attitudes Toward Statistics (SATS) (Gal, Ginsburg, & Schau, 1997). Additional items with regard to causal explanations were formulated using the same principles as the Attributional Style Questionnaire (ASQ) (Peterson et al., 1993), in particular on two attributional dimensions: stability and control. Finally, items were added to measure the two aspects of study behaviour: effort and persistence. All MSQ items concentrated on statistical related events. Because the MSQ was for the greater part based on existing surveys that have been proven to be valid (Peterson et al., 1993; Schau, Stevens, Dauphinee, & Del Vecchio, 1995), it can be considered an adequate measurement instrument regarding the relevant motivational constructs. Examples of questions are presented in Table 3.1. Based on content the items were divided into six subscales. To establish the reliability of the MSQ, it was administered to ($N$ = 200) first year Health Sciences students and Cronbach's alpha was computed for each subscale. Six questions that did not fit in the concerning subscale were identified. Four questions were removed, two were rephrased. The MSQ was the subsequent year used for collecting the data of the main study ($N$ = 94). It was administered to the students at the beginning of the introductory statistics course. Students received written instructions before they completed the MSQ. The whole procedure took approximately half an hour.

A second instrument was used to assess participants on effort and persistence, because it is well known that self reports and students' responses to questionnaires may not always adequately reveal mental processes, and behaviour (Biggs, 1993; Nisbett & Wilson, 1977; Schwartz, 1999; Watkins, 1996). The goal was to obtain more reliable data on study behaviour. The instrument consisted of two rating scales ranging from zero to ten. It was distributed to the tutors of tutorial group meetings. These are weekly two hour sessions supervised by a tutor, in which the students discuss the subject matter. The sessions are an essential part of the course. The tutors were given instructions on how to infer students' effort and persistence. They were told what was meant by effort and persistence, examples were given, and they were told how to use the rating scale (grades ranging from zero to ten are customary in our education). This came down to instructing them to ask and register whether students visited the lectures, whether students were prepared for the tutorial group meetings, and whether students were actively involved in the discussion during the obligatory meetings. The tutors had to convert their impression concerning these aspects

into a grade called effort. Persistence was analogously a grade based on the tutors' judgement concerning whether students kept on asking questions during the meetings until they really understood the subject matter, whether students at home persisted in trying to solve their assignments by using lecture notes and/or their books, or whether they consulted their teacher when they were not able to solve an assignment. The participants were, in the week before the end of the course, judged by their tutors. Finally, the scores on the exam at the end of the course were used as an indicator for participants' achievements. The exam consisted of 30 multiple choice questions and grades could range from zero to ten. Examples of questions of the exam are presented in the appendix.

*3.4.3 Analysis*

Sum scores of the responses to the questionnaire were computed for each subscale. Some items were positively phrased, others negatively. Responses on the negatively phrased items were mirrored so that all answers were in the same direction. These sum scores were called: *Stable Explanation, Control, Outcome Expectancy, Affect, Effort and Persistence.* To reflect the facts that people seek causes especially for failure (Graham & Weiner, 1987) and that motivation to study statistics usually is modest, the coding on the variables *Stable Explanation* and *Control* was done in such a way that high scores corresponded with respectively a stable negative explanation and lack of control. Cronbach's *α* was computed for each subscale. The exam grades (*Achievement*) and the tutor ratings *Effort(T)* and *Persistence(T)* consisted of grades ranging from zero to ten. They were included into the analyses as raw data.

Four analyses were done. First, several *t*-tests were done to test for possible selection biases. A comparison was done between the male and female participants on *Achievement, Stable Explanation, Control, Outcome Expectancy, Affect, Effort(T),* and *Persistence(T)*. Moreover, achievement was compared between the participants in our study and the rest of the cohort which took the introductory course. Second, bivariate correlations between all variables were calculated to inspect the correlation patterns. The covariance structure modelling was, because of the rather small sample size, done in two separate steps (Scott Long, 1983), resulting in the third and fourth analysis. The third analysis was a robust maximum likelihood confirmative factor analysis (the simultaneous analysis of the covariance and the asymptotic covariance matrix; Jöreskog & Sörbom, 1989), which was done to confirm the measurement structure. Fourth, a path analysis (a robust maximum likelihood structural equation modelling) was done with Lisrel. Due to the sample size it was necessary to disregard the measurement structure in this analysis. Hence, the analysis was done without latent variables, the sum scores of the separate items of the MSQ served as manifest variables. With this path analysis the model presented in Figure 3.1, was tested.

## 3.5 Results

To establish the reliability of the MSQ, Cronbach's α was computed in a pilot study ($N$ = 200) for each subscale. Six items with an insufficient fit were

identified. Four questions were removed, two were rephrased. Cronbach's α for each subscale (after the removal of the four items) and some example questions are presented in Table 3.1. The MSQ was then used for collecting the data of the main study in the subsequent year.

Table 3.1: Subscales of the MSQ

| *Subscales and example questions* | *Number of items* | *Cronbach's α* |
|---|---|---|
| *Stable explanation:* | 4 items | .8427 |
| Statistics is just a difficult subject. | | |
| I have always had difficulties with statistics. | | |
| *Control:* | 5 items | .7797 |
| The result on the statistics exam is determined by my own endeavour. | | |
| Whenever I don't understand a statistical topic, I know what to do. | | |
| *Outcome Expectancy:* | 6 items | .6048 |
| It pays off to study statistics. | | |
| The time I spend on statistics is wasted. | | |
| *Affect:* | 8 items | .7813 |
| To study statistics is enjoyable. | | |
| I think statistics is interesting. | | |
| *Effort:* | 8 items | .8058 |
| I spend a lot of time on statistics. | | |
| I never prepare myself for the statistics tutorial group meeting. | | |
| *Persistence:* | 7 items | .7405 |
| Whenever I don't understand something from statistics, I quit. | | |
| When I cannot complete a statistics assignment, I go through the book once again. | | |

Note: *N* = 200

A robust maximum likelihood confirmatory factor analysis was executed on those data of the MSQ that were also used in the path analysis of the main study (*N* = 94). The content based classification of the items on the subscales *Control, Stable Explanation*, *Outcome Expectancy,* and *Affect* was supported by the results of this confirmatory factor analysis; indices showed a proper fit. The Satorra-Bentler Chi square was used. It is considered to be more robust against a small sample size and violations of distributional assumptions (Hu, Bentler & Kano, 1992; Satorra & Bentler, 1994).

The Lisrel program provides several additional indices for how well the model fits the data (Jöreskog & Sörbom, 1988). A goodness of fit (GFI) is given for the whole model. It compares the tested model with a so called null-model, i.e. all parameters are fixed on zero. A second index is the normed fit index (NFI), which compares the tested model with an independence model (variances are set free, covariances are fixed on zero). This index, however, continues to improve when paths are added and therefore does not appraise parsimonious models adequately. The most meaningful index is the non-normed fit index (NNFI). In this index the degrees of freedom are taken into account and consequently it appraises not only the best fitting, but also the most parsimonious

model. All three fit indices should be close to one. Finally the root mean square residual (RMR) is given. This index should, as all residuals, be close to zero. The indices presented in Table 3.2 show a proper fit for this model, i.e. the items adequately fit into their subscales.

Table 3.2: Fit indices for the confirmatory factor analysis on *Control, Stable Explanation*, *Outcome Expectancy,* and *Affect*

| *Satorra-Bentler Chi square* | GFI | .86 | Standardised |
|---|---|---|---|
| (*df* = 224, *N* = 94) | NFI | .89 | RMR |
| | NNFI | .93 | |
| 277.18 (*p* = .04*) | CFI | .94 | .22 |

In Table 3.3 descriptive statistics of all the variables as measured by the MSQ, as well as the tutor ratings and the exam grades are given.

Table 3.3: Descriptives of the motivational variables and achievement

| | *M* | *SD* | *Number of items* | *Scale Min.* | *Scale Max.* | *Min. score* | *Max. score* | *Skewness* | *Kurtosis* |
|---|---|---|---|---|---|---|---|---|---|
| *Stable Explanation* | 16.39 | 5.61 | 4 | 4.00 | 28.00 | 4.00 | 28.00 | .100 | -.459 |
| *Control* | 16.22 | 4.84 | 5 | 5.00 | 35.00 | 5.00 | 31.00 | .697 | .637 |
| *Outcome Expectancy* | 29.13 | 5.13 | 6 | 6.00 | 42.00 | 14.00 | 40.00 | -.562 | .826 |
| *Affect* | 26.66 | 7.60 | 8 | 8.00 | 56.00 | 12.00 | 51.00 | .196 | .174 |
| *Effort* | 37.88 | 7.62 | 8 | 8.00 | 56.00 | 16.00 | 54.00 | -.477 | .347 |
| *Persistence* | 31.93 | 6.48 | 7 | 7.00 | 49.00 | 11.00 | 46.00 | -.207 | .293 |
| *Effort(T)* | 7.11 | 1.55 | 4 | 0.00 | 10.00 | 2.00 | 10.00 | -.803 | 1.212 |
| *Persistence(T)* | 6.64 | 1.73 | 4 | 0.00 | 10.00 | 1.00 | 10.00 | -1.028 | 1.862 |
| *Achievement* | 7.05 | 1.90 | 30 | 0.00 | 10.00 | 1.60 | 9.40 | -.780 | -.133 |

The results of the *t*-tests showed no significant differences between female and male participants. This might partly be because of the restricted power of the tests, so additionally the effect sizes (Cohen's *d*) were computed. The results are respectively for *Achievement* ($p$ = .65; $d$ = .13)*, Control* ($p$ = .51; $d$ = .17)*, Stable Explanation* ($p$ = .08; $d$ = .53)*, Outcome Expectancy* ($p$ = .49; $d$ = .22)*, Affect* ($p$ = .97; $d$ = .008)*, Effort(T)* ($p$ = .94; $d$ = .02)*,* and *Persistence(T)* ($p$ = .52; $d$ = .20).

Table 3.4: Correlations between the motivational variables and achievement.

| | *Stable Explanation* | *Control* | *Outcome Expectancy* | *Affect* | *Effort* | *Persistence* | *Effort(T)* | *Persistence(T)* | *Achievement* |
|---|---|---|---|---|---|---|---|---|---|
| *Stable Explanation* | 1 | .584* | -.336* | -.550* | .156 | -.052 | -.116 | -.138 | -.392* |
| | | $p$ = .000 | $p$ = .001 | $p$ = .000 | $p$ = .067 | $p$ = .310 | $p$ = .132 | $p$ = .093 | $p$ = .000 |
| *Control* | | 1 | -.647* | -.306 | .152 | -.099 | -.020 | .016 | -.121 |
| | | | $p$ = .000 | $p$ = .003 | $p$ = .072 | $p$ = .172 | $p$ = .423 | $p$ = .439 | $p$ = .123 |
| *Outcome* | | | 1 | .312* | .020 | .157 | .127 | -.006 | .226 |
| *Expectancy* | | | | $p$ = .001 | $p$ = .424 | $p$ = .065 | $p$ = .112 | $p$ = .479 | $p$ = .015 |
| *Affect* | | | | 1 | .125 | .239 | .266 | .216 | .429* |
| | | | | | $p$ = .115 | $p$ = .010 | $p$ = .005 | $p$ = .018 | $p$ = .000 |
| *Effort* | | | | | 1 | .746* | .273 | .263 | .261 |
| | | | | | | $p$ = .000 | $p$ = .004 | $p$ = .005 | $p$ = .006 |
| *Persistence* | | | | | | 1 | .368* | .337* | .294 |
| | | | | | | | $p$ = .000 | $p$ = .001 | $p$ = .002 |
| *Effort(T)* | | | | | | | 1 | .843* | .455* |
| | | | | | | | | $p$ = .000 | $p$ = .000 |
| *Persistence(T)* | | | | | | | | 1 | .478* |
| | | | | | | | | | $p$ = .000 |
| *Achievement* | | | | | | | | | 1 |

* $p \leq 0.001$ (Bonferroni corrected).

Combined these results indicate no substantial differences between male and female participants. An additional t-test was done to test for another possible selection bias. In this t-test the achievement of the students who participated in our study was compared to the rest of the cohort ($N$ = 122). No significant difference was found, nor a consequential effect size ($p$ = .82; $d$ = .06).

A correlation matrix of all variables was computed and is presented in Table 3.4. The significance level was adjusted with Bonferroni correction. Both dimensions of attribution (*Stable Explanation* and *Control*) were significantly correlated to *Outcome Expectation*. The notion of having no control was most strongly correlated to *Outcome Expectation*. *Outcome Expectation* was significantly correlated with *Affect* toward statistics.

*Affect* was significantly correlated to *Achievement*, but as expected not to the self reported behavioural constructs (*Effort* and *Persistence*), which were also not correlated to *Achievement*. The tutor ratings *Effort(T)* and *Persistence(T)* on the other hand were much better predictors for *Achievement* and were higher correlated to *Affect*. This is consistent with research that established the inaccuracy of self-reports and research that showed that students' responses to questionnaires may not always adequately reveal their own learning (Biggs, 1993; Glenberg, Sanocki, Epstein & Morris, 1987; Nisbett & Wilson, 1977; Schwartz, 1999; Watkins, 1996).

A path analysis with Lisrel was, because of this above-mentioned inaccuracy of self-reports, conducted on a model where the tutor ratings *Effort(T)* and *Persistence(T)* were inserted, instead of the self-reported study behaviour (*Effort* and *Persistence*). We started with our model that was presented in Figure 3.1. The relation between *Stable Explanation* and *Outcome Expectancy* based on attributional theories was not significant (Standardised Path coefficient β = .06; $p$ = .31). We did find a strong negative relation between the notion of having no control (*Control*) and *Outcome Expectancy* (β = -.68; $p$ < .000). Apparently, if a student thinks that there is no contingency between, for example, his study activities and the result on an exam, he will not expect a positive outcome of his actions.

The relations among the motivational constructs as well as the coefficients are displayed in Figure 3.2. The arrows in Figure 3.2 stand for the theoretical relations that were confirmed, the dotted arrows stand for the theoretical relations that were not confirmed, and the dashed arrows indicate meaningful relations that were not in the hypothesised theoretical model, as shown in Figure 3.1.

Figure 3.2 shows a strong direct relation between *Stable Explanation* and *Affect*, i.e. if students think that there are stable causes for negative statistics related events, failing their exams for example, they will develop negative feelings toward statistics. In the model, as displayed in Figure 3.1, this relation was mediated by *Outcome Expectancy*.

A negative *Outcome Expectancy* also had an adverse effect on *Affect*. *Affect* is related to all other constructs except to the notion of no control (*Control*) (β = .19; $p$ = .08). To emphasise the importance of *Affect*, it has been placed in a more central position in Figure 3.2. It is strongly related to *Achievement* directly,

as well as via *Persistence(T)*. Important is also that *Achievement* is determined by *Persistence(T)* (β = .34 ; *p* < .000) but not by *Effort(T)* (β = .08 ; *p* = .36).

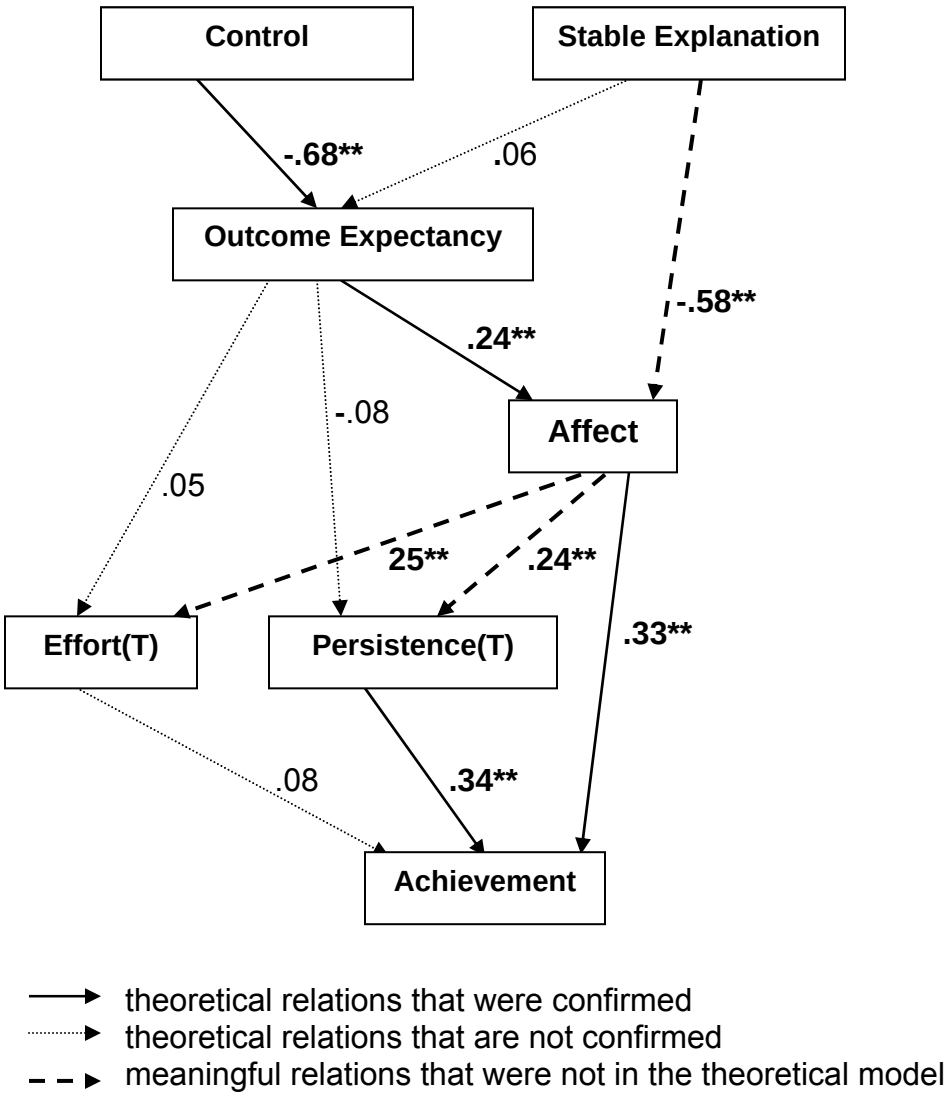



Figure 3.2. Statistics motivational model, as confirmed by path analysis (Lisrel).
Notes: Coefficients are standardised; * *p* < *0*.05; ** *p* < 0.01.

To enhance the fit of the model the residuals of the behavioural constructs *Effort(T)* and *Persistence(T)* had been set free to correlate (error covariance = 2.12; *t* = 5.34) in Lisrel. With this relaxation of the model (as presented in Figure 3.2), all fit indices showed a good fit. The values of these indices for our model are provided in Table 3.5. Again the Satorra-Bentler Chi square is presented, because of its robustness against a small sample size and violations of distributional assumptions (Hu, et al., 1992; Satorra & Bentler, 1994).

Table 3.5: Fit indices for the model in Figure 3.2

| | | | |
|---|---|---|---|
| *Satorra-Bentler Chi square* | GFI | .96 | Standardised |
| (*df* = 7, *N* = 94) | NFI | .95 | RMR |
| | NNFI | .93 | |
| 13.40; (*p* = .063) | CFI | .98 | .042 |

## 3.6 Discussion

This study was done in an introductory statistics course. It focussed on causal explanations of statistics related events, perceived outcome expectancy of students' activities within this statistics course, affect and study behaviour toward statistics, and the relation of these constructs to the results on the exam at the end of the course. These constructs were chosen because of their practical implications for the teaching of statistics.

Our first findings concern causal explanations. In the two presented motivational theories, perceived causes for events have underlying properties that have affective, behavioural, and cognitive consequences (Peterson, et al., 1993; Pintrich & Schunk, 1996). In our study we focussed on the dimensions control and stability of causal explanations.

The first result concerns control. The model in Figure 3.2 indicates that the perception of having no control over causes of statistics related events, may lead to decreased outcome expectancy. For example, a student who thinks that there is nothing he can do about the causes for failing the statistics exams, or thinks that he is not able to understand statistics anyway, may not expect a positive outcome of visiting the lectures or studying the material. This mechanism is intuitively appealing.

The second result indicates that the stability of causal explanations may be more directly related to affect. As is seen in Figure 3.2 we found a significant path from *Stable Explanation* of such causes to *Affect*. The path that we found may be interpreted as follows. The perception of stable causes for aversive events related to statistics, may lead to displeasure and frustration. If students perceive that failing statistics exams is not easily changeable, students may start to dislike statistics. This was reflected in responses like: *I dislike statistics; I do not have a positive perception of statistics, etc*.

In sum these two findings indicate that students who think that they lack control may not expect to profit from studying statistics, and students who do invest time, but think that there are stable causes for failing in spite of that, may start to dislike statistics.

The last path from *Stable Explanation* to *Affect*, though intuitively appealing, was not anticipated. The model in Figure 3.1 contained a relation between *Stable Explanations* and *Outcome Expectancy*. This relation was based on the general attributional position that the stability of a cause has the most influence on shifts in expectancy (Pintrich, 2000; Pintrich & Schunk, 1996; Weiner, 1986, 1992). Our findings are more consistent with the basic assumption from Peterson et al. (1993) that controllability is the major factor influencing

outcome expectancy. Yet, the direct influence of *Stable Explanation* on *Affect* may also have important practical implications for the education of statistics.

The implication for education from our findings may be that when students discover that the material is comprehensible to them and they experience success, they will be stimulated to study the material. This means that in constructing a learning environment, there should be tasks built in that are feasible for students. In that way the sequence of events that may lead to diminished motivation (Weiner, 1986) may be interrupted. Students will gradually sense that they can master the topics; they will discover they can control their learning outcomes, they will experience success, and they will abandon the idea that there are stable causes for failure. Control over learning outcomes may foster the positive expectation of future study activities. This positive expectation together with the reduction of the perception of stable negative causes for failure, may even promote students to enjoy studying statistics. Only then, should more difficult tasks be administered.

A second finding in our study is the central position of *Affect* in our model in Figure 3.2. Students who appreciate the value and relevance of statistics, who think it is interesting, challenging, and who like statistics, appear to study statistics more and qualitatively better, and perform better on the exams. In attributional theories (Pintrich, 2000; Pintrich & Schunk, 1996; Weiner, 1986, 1992) as well as in the learned helplessness theory of Peterson et al. (1993), affect is on the same level as behaviour and cognition. In the model in Figure 3.1 *Affect* was therefore put on a par with behavioural consequences of *Outcome Expectancy*. However, affect seems to have a more prominent role in motivational processes in the present statistics educational context. In our study we found that *Affect* directly and positively influenced *Achievement*. It also influenced study behaviour, i.e. *Effort(T)* and *Persistence(T)*. *Persistence(T)* on its turn also influenced *Achievement*. Thus, affect seems to determine achievement directly, as well as indirectly. Moreover, we found that *Affect* functioned as a mediator between *Control*, *Stable Explanations,* and *Outcome Expectancy* on the one hand, and the rest of the motivational constructs on the other. For this reason *Affect* holds a more central position in our model in Figure 3.2, than in the model which is presented in Figure 3.1.

The central role of *Affect* suggests that the students' feelings toward statistics appear to be an important theme for innovating and improving statistics education. Our results with respect to *Affect* are in line with Malone & Lepper (1987), who state that implementing features that make learning more appealing, enjoyable, and challenging makes learning more intrinsically motivating. Our finding that the feelings toward statistics are crucial in reaching satisfactory achievements also corroborates the results of Isen, Daubman & Gorgoglione (1987). In their study they found that positive affect may foster people's tendencies to see relations among stimuli, because positive affect leads to different ways of information processing, e.g. using different strategies. More relations between concepts are characteristic for richer knowledge networks, which indicate better integrated knowledge, and deeper understanding (Kintsch, 1988; 1998).

It seems, therefore, to be of relevance in the improvement of statistics education to make statistics courses more attractive, interesting, and enjoyable. One of the ways this might be achieved is by making the courses less theoretical. We think that a small experiment may engage students in a more active way, it may be fun to analyse data that are collected by the students themselves, and it may foster the notion of relevance of statistics.

A final result in our study was that *Effort(T)* had no significant relation with *Achievement*. Both *Effort(T)* and *Persistence(T)* were determined by the tutors. *Effort(T)* mirrored the amount of time that was studied, and the fact whether students prepared themselves, visited lectures, or were actively involved in the discussion during group meetings. Effort per se seemed to have a minor effect on achievement. What counts seems to be the way students study. In our study, *Persistence(T)* contributes significantly to the result on the exam. Students who did not quit that easily, who persisted, who turned to their lecture notes or their books, or consulted a teacher when they were not able to solve a statistical problem, those students did better on the exam. This result suggests that persisting is the best way to study statistics. It is in line with research in other subjects that established the importance of learning strategies and mastery goals for achievement in educational settings. (see e.g., Ames, 1992; Boekaerts; 1997; Pintrich, 2000; Dweck, 2000).

This finding may also be important for educational purposes. In the teaching of statistics, students should be stimulated to try to solve their problems. They should try to persist instead of quitting all too easily. This can be done by guiding them through the topics and by pointing them into the correct direction, instead of giving the solution to a problem promptly. Persisting and learning from mastering their own difficulties may be the most valuable way of learning.

The student population from which we recruited our participants consists largely of female students. Consequently, most of our participants were female (79 female versus 15 male). This could have affected our results. However, *t*-tests on all the core variables (*Control, Stable Explanation, Outcome Expectancy*, *Affect, Effort(T)*, *Persistence(T),* and *Achievement*) in our models showed no significant differences between the female and male students. Therefore, the fact that the majority of our participants was female seems not to affect the motivational processes that were studied.

The tutor ratings that we used to measure effort and persistence are another limitation of our study. We instructed the tutors in great detail and asked them to register during the course students' activities that we hold indicative for effort and persistence. We are confident that the ratings of the tutors are a quite valid and reliable measurement of the relevant behaviour. Still these ratings only reflect observable, external behaviour. Consequently we cannot discuss internal processes of reflection and mental activity. Our results only pertain to self-reported cognitions, affect, and observed behaviour.

In the present study only first year students were studied. In future research second and third year students could be studied. Secondly, our results could be corroborated in studies with a larger sample. In our study a rather small sample was used ($N$ = 94). Thirdly, it could be investigated how in a practical

educational context it can be realised that students persist during studying statistics. How can students optimally be guided to the correct solution of the problems? Will this reduce the perception of stable negative causes for failure and enhance the notion of control? Will such a reduction lead to a positive expectation of future study activities and to more enjoyment? Will all this eventually lead to more persistence and better results on the exam? Finally, further research is needed to investigate what the most effective way is of making statistics education more enjoyable.

### 3.7 Appendix

Example questions from the exam at the end of course (used for the measurement of achievement).

1. In a sample of 101 newborn babies, the mean birth weight is 3.8 kg and the standard deviation is 0.85. The null hypothesis is $H_0$: μ = 4 kg.
If this null hypothesis holds, then:
a. The probability that we will find a sample mean smaller than or equal to 3.8 kg is 50%
b. The probability that we will find a sample mean smaller than or equal to 3.8 kg is 80%
c. The probability that we will find a sample mean smaller than or equal to 3.8 kg is less than 50%
d. The probability that we will find a sample mean smaller than or equal to 3.8 kg is greater than 50%

2. Given the same sample as in question 1, we are testing $H_0$: μ = 4 kg against $H_1$: μ ≠ 4 kg. The *p*-value of the sample mean of 3.8 kg is:
a. $p \leq .01$
b. $.01 < p \leq .02$
c. $.02 < p \leq .05$
d. $p > .05$

3. Given the same sample as in question 1, we are again testing $H_0$: μ = 4 kg against $H_1$: μ ≠ 4 kg. Suppose the null hypothesis is rejected at $\alpha = .10$. What is the implication of this $\alpha = .10$?
a. In 10 % we will wrongfully conclude that $H_0$: μ = 4 kg holds.
b. In 10 % we will wrongfully conclude that $H_1$: μ ≠ 4 kg holds.
c. In 5 % we will wrongfully conclude that $H_0$: μ = 4 kg holds.
d. In 5 % we will wrongfully conclude that $H_1$: μ ≠ 4 kg holds.

4. The effects of 3 instructional methods on comprehensibility of the information (SCORE) were investigated. The 3 methods were: a standard method and 2 experimental methods (experimental method 1 and experimental method 2). The coding of the dummy variables was as follows:

| | D_EXP1 | D_EXP2 |
|---|---|---|
| Standard method | 0 | 0 |
| Experimental method 1 | 1 | 0 |
| Experimental method 2 | 0 | 1 |

It is tested whether the comprehensibility of the information (SCORE) for all methods is equal ($H_0$), or if at least one of the three methods is different ($H_1$).

Part of the output of the SPSS analysis is presented below:

ANOVA

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 3341.722 | 2 | 1670.861 | 6.317 | .005 |
| | Residual | 8727.917 | 33 | 264.482 | | |
| | Total | 12069.639 | 35 | | | |

Predictors: (Constant), D_EXP2, D_EXP1
Dependent Variable: SCORE

Coefficients

| | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| Model | | B | Std. Error | Beta | | |
| 1 | (Constant) | 37.750 | 4.695 | | 8.041 | .000 |
| | D_EXP1 | 14.250 | 6.639 | .367 | 2.146 | .039 |
| | D_EXP2 | 23.417 | 6.639 | .603 | 3.527 | .001 |

Dependent Variable: SCORE

Question: What conclusion can be drawn? Assume $\alpha = 0.05$.

a. There is a difference between the instructional methods because the *p*-value of *F* is smaller than 0.05.
b. There is a difference between the instructional methods because the *p*-value of *F* is smaller than 0.05/2 =0.025.
c. There is no difference between the instructional methods because the *p*-value of *F* is smaller than 0.05/2 =0.025.
d. There is no difference between the instructional methods because the *p*-value of *F* is smaller than 0.05.

Given the same research and the same results as in question 4, suppose that the F-test indicates a difference between the three methods. Which groups differ significantly? Assume $\alpha = 0.01$.

a. Each method differs significantly from the others.
b. The standard method differs significantly from experimental method 1.
c. The standard method differs significantly from experimental method 2.
d. The experimental method 1 differs significantly from experimental method 2.

5. Given the same research and the same results as in question 4, what is the proportion of explained variance in the SCORE variable?

a. 0.28
b. 0.72
c. 0.38
d. 0.86

# Chapter 4

## The effect of directive tutor guidance in problem-based learning of statistics on students' perceptions and achievement

This chapter is accepted for publication as: The effect of directive tutor guidance in problem-based learning of statistics on students' perceptions and achievement, Luc Budé, Tjaart Imbos, Margaretha W.J. van de Wiel, Martijn P.F. Berger (2007). Higher Education.

## 4.1 Abstract

*In this study directive tutor guidance in problem-based learning (PBL) of statistics is investigated. In a quasi experiment in an educational setting, directive guiding tutors were compared with tutors in a more traditional role. Results showed that the subjective perceptions of the students with regard to the course, the tutor, and the discussions in the tutorial meetings were more positive in the guided condition. The quality of the problems used in the meetings and general tutor functioning were evaluated as equal in both conditions. Achievement was marginally higher in the guided condition. It can be concluded that directive tutor guidance is an effective addition to PBL of statistics.*

## 4.2 Introduction

In this study the role of the tutor in problem-based learning (PBL) of statistics is investigated. The benefit of human tutoring has been demonstrated in many studies (Chi, 1996, Chi, Siler, Jeong, Yamauchi, & Hausman, 2001; Graesser, Person, & Magliano, 1995). Positive aspects of tutor performance in problem-based learning (PBL) have been studied extensively too (Barrows, 1988; Dolmans, et al., 2002; Dolmans & Wolfhagen, 2005; Schmidt & Moust, 1995). Summarising these findings it can be concluded that tutoring is effective:

1. When tutors and students interact; tutor's actions that prompt for the co-construction of knowledge are positively correlated with deep understanding (Chi, 1996, Chi, et al., 2001)
2. When students are activated; understanding will be improved by answering why, how, and what-if questions (Graesser, Bagget, & Williams, 1996)
3. When self-study is stimulated; self-study will induce higher achievement (Schmidt & Moust, 1995, 2000).

One-on one tutoring can be defined as a way of instruction, which is characterised by an interactive and continuous stream of exchanges between a tutor and a tutee (Chi, 1996). Although tutors in PBL operate in a group, their role is more or less the same as in one-on one tutoring. Instead of dispensing knowledge they should try to activate students, stimulate group processes, try to create an atmosphere in which students can optimally participate in the discussions, help students to monitor their own learning, and to stimulate self-study (Schmidt & Moust 2000; De Grave, Dolmans, & Van Der Vleuten, 1999).

In PBL both content expert faculty tutors and non-content student tutors are used. Content expert tutors tend to make more subject matter contributions to the discussions than non-content expert tutors. This has been shown to improve especially novice student's performance. It seems that a tutor's expertise can compensate for lack of prior knowledge of students (Schmidt, v.d. Arend, Moust, Kokx, Boon, 1993, Dolmans, et al., 2002). Research has also shown that tutorial groups with relatively low levels of productivity require more input from a tutor (Dolmans & Wolfhagen, 2005). To be able to make subject matter contributions the tutors in PBL usually have tutoring instructions at their disposal. These instructions consist of general information about the course, the

main topics, more specific information about the goal of the problems, which subject matter is supposed to be discussed and how to tackle the problems. In the intervention of the current study the tendency of tutors to intervene in the content of the discussion, was taken a step further.

The intervention consisted of additional tutor guidance during the meetings. That is, the tutors actively guided the discussions in a more directive way than it is usual in PBL. The research question of our study was concerned with the effect of this extra directive tutor guidance on students' achievement and their subjective perception of several course aspects, group functioning, the relevancy of tutor contributions, general tutor functioning, and the quality of the problems that were used for the discussions in the tutorial sessions.

In the following we will first give a short description of PBL. Next, we will describe some characteristics of statistical knowledge in relation to the learning of statistics and we will describe our intervention. After that, we will present our study.

### 4.3 Problem-based learning

PBL refers to a variety of approaches to instruction, which all have in common that much of the learning and instruction is anchored in concrete problems (Hmelo & Evensen, 2000). A problem can be anything that raises questions germane to the subject matter and affords free inquiry by students (Barrows, 1986). For example, a problem can be a patients' case (in medical education), the outcome of a study, a hypothesis about a real life phenomenon, or a statistical problem that students need to solve.

While exploring and discussing the problem in an initial tutorial session, students extract key information about the problem and discover deficiencies in their knowledge (Hmelo & Evensen, 2000). Based on these knowledge gaps and out of an intrinsic curiosity, students formulate their own learning goals and decide on what they are going to study. Instead of merely being exposed to information, the students in PBL are actively engaged in gathering information that is processed in relation to the presented problem. After the initial tutorial session the students in most PBL courses individually study relevant literature and in the following session they report back to the group what they have learned.

During the discussions in the tutorial sessions there are two kinds of guidance. Firstly, the problems used in PBL usually provide a fair amount of guidance to the students. Ideally, a problem directs students to the main topics and focuses on important issues. As a consequence the effectiveness of PBL is closely related to the quality of the problems (Albanese & Mitchel, 1993; Schmidt & Moust, 2000).

Secondly, effective tutoring, as explained in the introduction, also involves appropriate guidance from the tutors in the discussions of the subject matter. This guidance usually consists of hinting the students, pointing to relevant topics, and helping students to monitor their own learning. Tutors traditionally ask questions like: can you explain this; do you understand that; can you see why

that is important, etc? However, if a tutor contributes too much to the discussion in PBL, self-study time decreases (Schmidt & Moust, 2000).

Moreover, research has shown that education in general should not be too directive. Otherwise, students may loose the idea of having control, which may cause self-regulated learning to decrease, which on its turn may cause students to get less motivated and become more passive (Ames, 1992; Deci & Ryan, 1985; Lepper, Drake, & O'Donnell-Johnson, 1997; Pintrich, 2003). Passiveness is unfavourable, because the best way to acquire knowledge is to actively process the subject matter (see e.g., Anderson, 1983; Graesser, et al., 1996; Chi, Bassok, Lewis, Reiman, Glaser, 1989).

In spite of these findings we instructed the tutors in the intervention condition of our study to guide the discussions in a directive way. We think that in statistics such an approach might have beneficial effects, as will be explained in the next paragraph.

### 4.4 Characteristics of Statistical Knowledge

Statistics is difficult to teach and often poorly understood (Gal & Garfield, 1997). Probably this is due to the fact that statistical concepts are relatively abstract and interconnected to a great degree (Schau & Mattern, 1997). For example, the fact that the mean is sensitive to outliers can affect the interpretation of the results of an analysis of variance. This example also illustrates that statistical knowledge is hierarchical. A student first has to know what the mean is, than has to understand the concept deviation from the mean, before variance can be understood, which has to be comprehended before analysis of variance can be understood. Statistical knowledge is said to be highly structured, because of this interconnectivity of the concepts and the hierarchy of the knowledge.

In abstract and highly structured knowledge domains (statistics, computer programming, and mathematics) research has shown that learning can be improved when teaching is more directive. For example, by using worked examples in mathematics (Carroll, 1994; Sweller, 1999; Sweller & Cooper, 1985), and computer programming (Tuovinen & Sweller, 1999); guided discovery in computer programming (Debowski, Wood, & Bandura, 2001; Fay & Mayer, 1994; Lee & Thompson, 1997); and by explicating strategies in statistics (Paas, 1992). The interventions in all these studies acted as a guide, showing the students the most effective way to achieve their aim. In all studies it was shown that the performance of students improved. Our intervention intended to have the same effect.

### 4.5 Directive tutor guidance

As explained in the paragraph on PBL, students in PBL are directed by the problems and the tutors. It was expected that for statistics education extra directive guidance provided by the tutor would promote students' understanding and that students would gain a better insight into the content of the course. In the current study the tutors in our intervention condition provided directive guidance by asking questions. They received a detailed list of specific questions, in

addition to the more general tutoring instructions with background information that tutors usually have. By asking questions, the tutors in the intervention condition initiated the discussion and they directed the discussion in a predetermined way. The tutors intervened as soon as the students would go astray or omit an important subject.

The questions were based on how experts would deal with the presented problems. Step by step, in a subscribed order the tutoring questions guided the students through the subject matter. The guiding questions focussed the discussions on relevant issues, indicating the direction of reasoning without hampering the active learning processes. Tutors received written directions and training in how to use the instructions. They were instructed to specifically ask for the relations between the concepts. This should make more explicit the interconnectivity of the concepts. The tutors were also instructed to make sure that the topics were discussed in a prescribed order. This was done to adhere to the hierarchy of the knowledge. We will refer to this intervention as directive tutor guidance.

It was hypothesised that this directive tutor guidance would have a positive effect on students' achievement and subjective perception of the course, group functioning, and the relevancy of tutor contributions. No effect was anticipated with regard to the subjective perception of general tutor functioning and the quality of the problems that were used for the discussions in the tutorial sessions.

## 4.6 Method

### *4.6.1 Participants*

Two hundred and six students, enrolled in a bachelor statistics course of Health Sciences at the University of Maastricht, participated in this study. They were randomly assigned to 24 tutorial groups. These 24 tutorial groups were randomly assigned to 14 tutors (ten tutors had two groups, four had one group). Finally, the tutors were assigned to either the guided condition or to the control condition using blocked randomisation. This resulted in 12 groups (N=102) being assigned to the guided condition and 12 groups (N=104) to the control condition. The tutors were members from the department of Epidemiology and the department of Methodology and Statistics. All tutors had sufficient knowledge of the subject matter and had minimal three years experience in tutoring.

### *4.6.2 Materials and procedure*

The bachelor statistics course, in which this study was conducted, took eight weeks. Students discussed the topics in a weekly tutorial group meeting on the basis of a list of problems. The main topics were methodological and statistical subjects. The methodological topics were randomised clinical trials and quasi experimental designs. The statistical topics were the central limit theorem, *t*-tests, ANOVA, and linear regression analysis. For an example of a problem see appendix A.

In the control condition, tutors followed the usual PBL tutoring procedure, facilitating the group and learning processes, but without directing the

discussions. In the intervention condition, the tutors used a list of questions to guide the discussion in the tutorial group meetings in a directive way. The list consisted of questions for each statistical problem that was discussed in the tutorial meetings. See appendix B for an example.

The questions were in a prescribed order, oriented toward the main topics and their relations. Typically, in PBL the discussion starts with the problem definition. The tutor would therefore start with the question what the problem was about and successively guide the students through the subject matter. The questions were used in the initial discussion to make sure that all relevant topics were covered in the correct order and students would see what they had to know about the subject matter and what they did not yet know. This enabled the formulation of correct learning goals. After individual study, in the reporting phase, the questions were used during the discussions to try to stimulate that students would grasp all the concepts and their relations.

Tutors received training in handling the tutor instructions. It was stressed to only ask questions to keep the students actively engaged and not to explain the issues that were being discussed. Moreover, tutors were instructed to ask a question only when students were in danger of wandering off, when omissions were made, or when students did not know how to proceed. Finally, they were instructed to specifically emphasise the relations between the discussed concepts. For example, they were supposed to ask how sample size is related to the standard error.

At the end of the course the students evaluated the course on various aspects in a questionnaire. All courses are evaluated as a standard procedure. This is why students are used to filling out evaluation questionnaires after the final course exam. The questionnaire used in this study had 19 items consisting of statements that students had to rate on a five point Likert-scale. The statements covered the course itself, the statistical problems used in the meetings, the tutorial meetings, the performance of the tutor in stimulating understanding, and three more traditional aspects of the general functioning of the tutor, with respect to facilitation of the group and learning processes. The statements are presented in appendix C. To assess students' achievement, the final course exam was used. This exam consisted of 30 multiple choice questions about the statistical and methodological subject matter of the course.

*4.6.3 Analysis*

The raw scores of the questionnaire and the final course exam scores were used for the analysis. The item scores were inspected with respect to the mode, skewness and kurtosis. A factor analysis was done on the individual items to distinguish the subscales. This resulted in five subscales, making up five dependent variables. For each variable Cronbach's alpha was computed.

This study had a hierarchical design. First, the students were randomly assigned to the tutorial groups. Next, tutors were randomly assigned to the guided and the control condition of this study. Because of this hierarchical design, the comparison between the two conditions, with respect to the five dependent variables, was done by means of multi-level analyses. Random

intercept regression models were used for all five analyses, with the students as the first level and the tutors as the second level. Deviance tests were used for the random effects, because of the rather small sample size (Snijders & Bosker, 2003).

## 4.7 Results

### *4.7.1 The evaluation questionnaire*

Inspection of the item scores with respect to the mode, skewness and kurtosis showed no indication of violation of the normality assumption. Factor analysis resulted in two possible solutions. The scree plot indicated a three factor solution (see Figure 4.1).

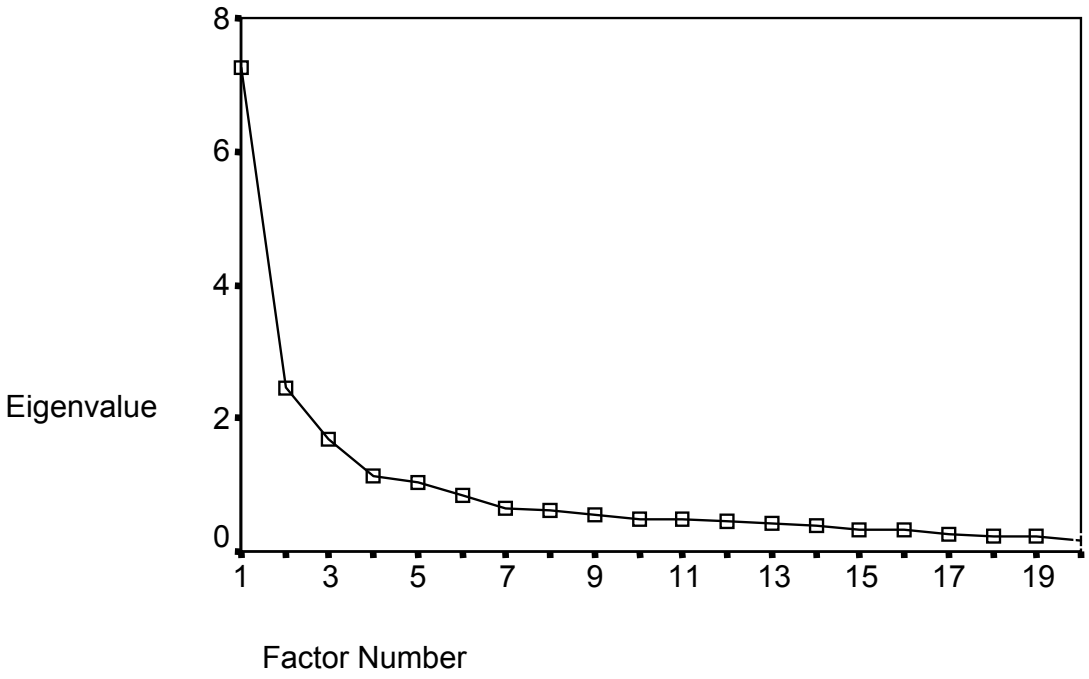



Figure 4.1. Scree plot of the factor analysis on the individual items

The “eigenvalues greater than one criterion” indicated five factors. The eigenvalues for the five factors ranged from 7.00 to 1.02. We have chosen for the five factor solution, because after oblique rotation a clear pattern emerged that could directly be interpreted in relation to the content of the resulting five subscales (see appendix C). The pattern matrix after oblique rotation is shown in Table 4.1. The highest factor loadings are in bold type.

Table 4.1. *Results of factor analysis on the items of the questionnaire after Oblimin rotation*

| | Factor | | | | |
|---|---|---|---|---|---|
| | *Course* | *Problems* | *Elaboration* | *Tutor guidance* | *Tutorgeneral* |
| 1. Clarity | **.725** | .048 | -.100 | -.009 | -.101 |
| 2. Instructiveness | **.779** | -.002 | .147 | .007 | .078 |
| 3. Organisation | **.560** | .010 | -.030 | .010 | -.073 |
| 4. Variation | -.003 | **.737** | -.050 | -.087 | .074 |
| 5. Systematic approach | .103 | **.404** | -.128 | .025 | .128 |
| 6. Useful learning goals | .041 | .193 | **-.613** | .012 | -.141 |
| 7. Useful discussions | .083 | -.026 | **-.613** | -.033 | .175 |
| 8. Effectiveness | -.041 | .105 | **-.777** | .031 | -.021 |
| 9. Integration | .105 | .105 | -.297 | -.304 | .301 |
| 10. Pleasant | .250 | -.203 | **-.442** | -.089 | .311 |
| 11. Productive | .097 | -.124 | **-.522** | -.185 | .319 |
| 12. Tutor systematically | -.002 | -.014 | -.187 | **-.596** | .256 |
| 13. Tutor insight | -.035 | -.022 | -.118 | **-.754** | -.047 |
| 14. Tutor structuring | .043 | -.037 | -.117 | **-.565** | .264 |
| 15. Expertise use | -.018 | .037 | .164 | **-1.001*** | -.131 |
| 16. Contributions relevant | .065 | .058 | .095 | **-.851** | .045 |
| 17. Autonomy stimulation | .059 | .016 | -.148 | -.248 | **.540** |
| 18. Activation | -.006 | .042 | -.183 | -.133 | **.519** |
| 19. Evaluation | -.052 | .161 | .087 | .021 | **.603** |
| | | | | | |
| Factor correlations | | | | | |
| Course | 1.000 | .103 | -.389 | -.230 | .165 |
| Problems | | 1.000 | -.291 | -.087 | .212 |
| Elaboration | | | 1.000 | .317 | -.409 |
| Tutorguidance | | | | 1.000 | -.482 |
| Tutorgeneral | | | | | 1.000 |
| | | | | | |
| Cronbach's α | .72 | .60 | .85 | .90 | .73 |

*Note*. Principal Axis Factoring with Oblimin rotation; * value larger than one, artefact due to oblique rotation.

Items that related to the course itself loaded highly on the first factor (items 1-3). The sum score of these three items comprises the variable *course*.

The first two items that related to the problems (items 4-5) loaded highly on the second factor, their sum comprises the variable *problems*.

The other two items about the problems (items 6-7; referring to what the problems added to the discussions) together with three questions with respect to the discussions in the tutorial meetings (items 8, 10-11) loaded highly on the third factor. One item referring to these discussions (item 9) loaded equally on three different factors. Based on the content of the item it was categorised under the third factor. The sum of these six items (items 6-11) comprises the variable *elaboration*.

Five items loaded highly on the fourth factor (items 12-16). These items were related to those aspects of the tutor functioning that were supposed to stimulate students' understanding of the subject matter. The sum of these five items forms the variable *tutor guidance*.

Finally, the three items with respect to the general functioning of the tutor (items 17-19) loaded highly on the fifth factor. Their sum forms the variable *tutor general*. Together the five factors explained almost 68 percent of the variance. Cronbach's *α* for the five variables was high: *course* (*α* = .72), *problems* (*α* = .60), *elaboration* (*α* = .85), *tutor guidance* (*α* = .90), *tutor general* (*α* = .73).

*4.7.2 Fixed and random effects*

The mean grade on the final course exam was higher for the students in the guided condition as hypothesised, although the difference was only marginally significant ($M_{control}$ = 5.85, $M_{guided}$ = 6.20; $t$ = 1.47; $p$ = .072). The means of the five variables for the two conditions are presented in Table 4.2.

Table 4.2. *Means of the five dependent variables for the guided and control condition*

| | *Number of items* | *Minimum* | *Maximum* | *Scale centre* | $M_{control}$ | $M_{guided}$ |
|---|---|---|---|---|---|---|
| *Course* | 3 | 3 | 15 | 9 | 8.38 | 9.34 |
| *Elaboration* | 6 | 6 | 30 | 18 | 15.70 | 18.19 |
| *Tutor guidance* | 5 | 5 | 25 | 15 | 16.14 | 19.78 |
| *Problems* | 2 | 2 | 10 | 6 | 4.52 | 4.47 |
| *Tutor general* | 3 | 3 | 15 | 9 | 7.80 | 8.68 |

The results of the multilevel analyses showed that in the guided condition the means of *course*, *elaboration*, and *tutor guidance* were significantly higher than in the control condition. No differences were found regarding *problems* and *tutor general*. The difference at the tutor level was not significant in the analysis of *course*. In the other four analyses these differences were significant. The results are presented in Table 4.3.

Table 4.3. *Random intercept multi-level analyses on the five dependent variables with tutor as the second level*

| | *b* | *S.E.* | *t* | *p*-value | | | *p*-value |
|---|---|---|---|---|---|---|---|
| *Course* | | | | | | | |
| Fixed Effects | | | | | Random effects | | |
| Intercept | 10.32 | .63 | 16.29 | .000* | $\tau^2_{tutor}$ | .11 | |
| Condition | -.99 | .40 | -2.48 | .028* | ΔDeviance | .37 | .270 |
| *Elaboration* | | | | | | | |
| Fixed Effects | | | | | Random effects | | |
| Intercept | 15.83 | 1.45 | 10.90 | .000* | $\tau^2_{tutor}$ | 1.95 | |
| Condition | -2.65 | .89 | -2.97 | .012* | ΔDeviance | 14.40 | .000* |
| *Tutor Guidance* | | | | | | | |
| Fixed Effects | | | | | Random effects | | |
| Intercept | 23.92 | 1.97 | 12.12 | .000* | $\tau^2_{tutor}$ | 4.10 | |
| Condition | -4.11 | 1.21 | -3.40 | .006* | ΔDeviance | 25.70 | .000* |
| *Problems* | | | | | | | |
| Fixed Effects | | | | | Random effects | | |
| Intercept | 4.46 | .02 | 7.19 | .000* | $\tau^2_{tutor}$ | .35 | |
| Condition | .62 | .38 | .052 | .959 | ΔDeviance | 13.60 | .000* |
| *Tutor General* | | | | | | | |
| Fixed Effects | | | | | Random effects | | |
| Intercept | 9.75 | 1.08 | 9.06 | .000* | $\tau^2_{tutor}$ | 1.13 | |
| Condition | -1.06 | .66 | -1.60 | .138 | ΔDeviance | 15.00 | .000* |

*Note*. * $p < .05$; ΔDeviance is the difference between the deviance of the model with and without the random tutor effect

## 4.8 Discussion

In this study, it was examined whether directive tutor guidance in problem based learning of statistics improved the subjective perceptions of the students regarding the course, the tutor, the discussions in the tutorial meetings, the quality of the problems used in the meetings, and general tutor functioning. Directive tutor guidance aimed at stimulating the students to actively link together the statistical concepts in a more structured way. It was expected that this would increase students' understanding of the topics and that students would gain a better insight into the content of the course.

The final course exam served as a measurement of achievement. The grades of the exam were used as an indication of students' understanding. We expected all students to be able to pass the exam at the end of the course, but

we expected the students in the guided condition to do better. The results confirmed our hypotheses. We found a marginally significant effect. Moreover, it should be noted that the difference between the conditions is relevant in practice. Students pass the exam if they get a grade of 5.5 (out of 10 points). The means of both conditions were above the cut off score. It should be noted, however, that the difference of 1.5 times the standard error in favour of the guided condition is very close to this cut off score. For a small group of students the positive effect of our intervention has been crucial. In the guided condition 67 percent of the students passed the final course exam, while 60 percent passed in the control condition.

With regard to the questionnaire the results showed that all the hypothesised effects occurred. The course was valued more positively by the students in the guided condition. These students also rated the discussions in the tutorial meetings higher. Those aspects of the tutor functioning that were supposed to stimulate students' understanding of the subject matter were also better evaluated in the guided condition. In both conditions the problems were judged of equal quality. Furthermore, general functioning of the tutors was judged as similar in both conditions.

The items of the questionnaire referring to the course consisted of statements regarding the clearness of the goal of the course, the instructiveness of the course, and the organisation of the course. Students in the guided condition evaluated the course more positively than the students in the control condition. The more positive evaluation indicates that the students in the guided condition had better insight in the content of the course and had a more positive overall impression, as hypothesised. This result suggests that students apparently understood better what was expected from them with respect to the objectives of the course, what and how they had to learn, and what they did learn was clearer to them. All statements in the evaluation questionnaire had to be rated by the students on a five point Likert-scale. This means that a neutral opinion with respect to the course would have resulted in a mean score of nine, i.e. nine is the centre of the scale. The means of the two conditions show that students in the guided condition were slightly above this centre of the scale (i.e. a positive rating), where students in the control condition rated the course slightly negative.

Items constituting the variable elaboration included statements concerning the integration of the subject matter in the discussions and the productiveness of the meetings. Items referring to particularly those aspects of tutor behaviour that were supposed to stimulate students' understanding constituted the variable tutor guidance. These items included statements like: the tutor helped structuring the subject matter and the tutor's contributions were relevant. The higher ratings of the students in the guided condition of both elaboration and tutor guidance indicate that a more directive tutor behaviour also had a positive effect on the instructiveness of the discussions. A neutral stance toward elaboration would have resulted in a mean score of 18. The means for the variable elaboration show that students in the guided condition rated the discussions positively, the students in the control condition negatively. Tutor guidance was judged positively

in both conditions, as students in both conditions rated tutor guidance above 15, the centre of the scale.

Directive guidance in education may not only have positive effects, but it may also lead to a decrease of self-study time and self-regulated learning, so students may become less motivated and more passive (Ames, 1992; Deci & Ryan, 1985; Lepper, et al., 1997; Pintrich, 2003; Schmidt & Moust, 2000). Therefore, directive tutoring might have led to a more negative evaluation of the category of items referring to the general functioning of the tutors. These items consisted of statements regarding the stimulation of autonomy, the activation of the students, and the evaluation of the group processes. However, no differences were found between the two conditions. It can be concluded that directive tutor guidance did not have a negative effect in this respect.

In both conditions the same problems were used to initiate the discussion in the tutorial group meetings. Therefore, no differences between the two conditions were hypothesised in students' judgments about the problems. This is exactly what we found. These findings together with the equal evaluation of the general functioning of the tutor in both conditions supported the other findings.

The results also show (except for the variable course) that tutors differ significantly from each other. The influence of the tutors on students' perception of the courses is relatively small. Their influence on the discussions and their own functioning is obviously much bigger. Multi-level analyses of those variables that are influenced by the tutors showed differences between tutors. For the course no differences were found. These results may seem trivial. However, for the uniformity of education it is important to try to increase similarity in the way tutors interact with the students. Our intervention exactly tried to do this. Tutors received specific questions and received training in how to use the instructions. This may have reduced the individual differences in tutoring style. As we did not directly observe tutor behaviour in the groups, but inferred this from students' responses on the evaluation questionnaire we do not know how the instructions were carried out in practice. Tutors who guide the discussion, as in our intervention, are more prominent in the meetings. As a consequence, directive guiding tutors might be inclined to explain some of the topics, although they were specifically instructed not to do so. Future research could be aimed at how the instructions influence differences between tutors and tutor behaviour per se in practice.

Future research could also be directed at the underlying mechanisms for the results that we found. Our results are in line with cognitive load theory. Providing guidance may have reduced extraneous cognitive load (van Merriënboer & Sweller, 2005; Paas, Renkl & Sweller, 2004). Asking specifically for explanations of the relations between the statistical concepts may have increased germane cognitive load (Paas & van Merriënboer, 1994; Sweller, van Merriënboer & Paas, 1998). It could be measured whether our approach has such an effect on the perceived cognitive load.

Finally, it is unclear which kind of students have profited most from the extra guidance. We assume that specifically poor students who have had

difficulty in mastering the subject matter, may have had the most benefit from the extra guidance. Future research is needed to confirm this.

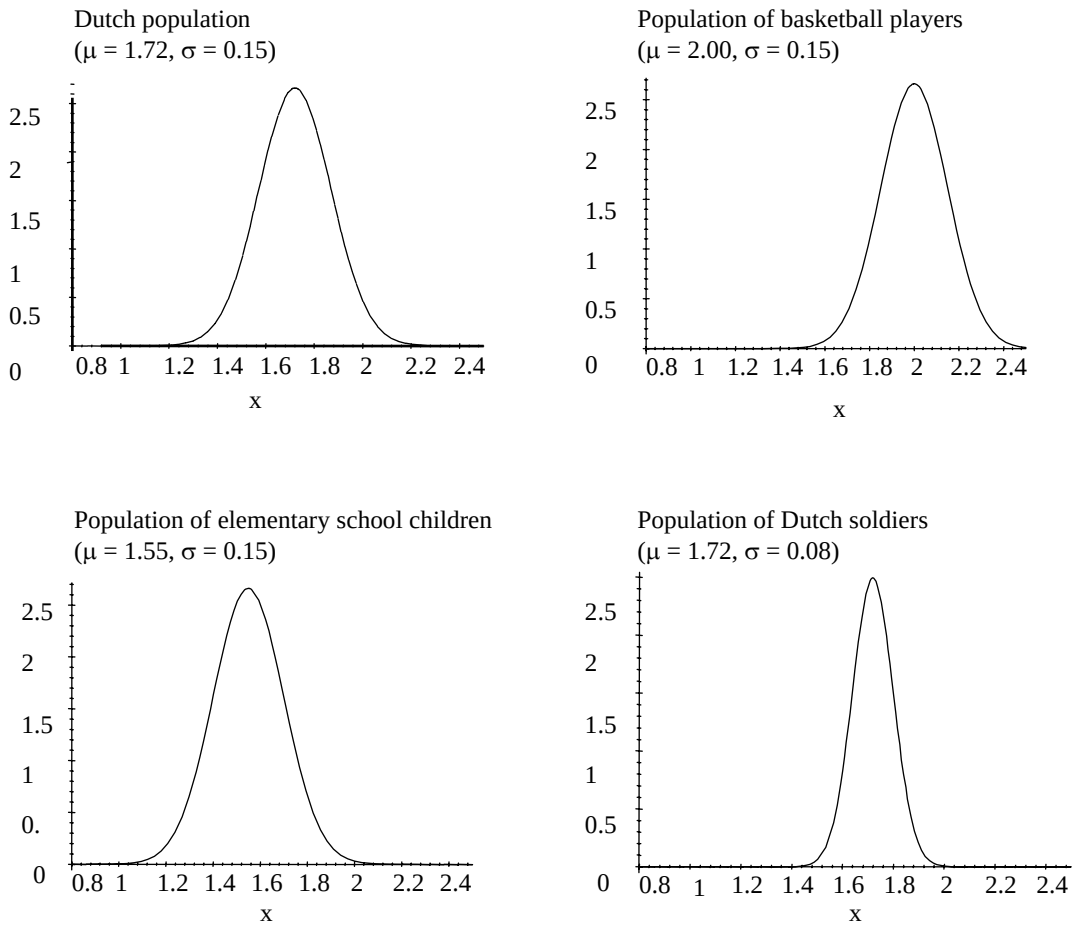
This study was done in a field setting. On the one hand this limits the scope of our conclusions. We used a standard questionnaire and the regular final course exam as measurements. Moreover, we could not control all factors. For example, we do not have a complete view of how tutors behaved in the meetings, nor do we know how students studied in the different conditions. On the other hand, because the studied conditions were embedded in realistic learning situations, we think that the outcomes are relevant in practice.

## 4.10 Appendix A

### Example of a PBL problem for statistics.

The presented problem is an introduction to the idea of sampling distributions. The problem tries to raise the notion of the three aspects (population mean $\mu$, population standard deviation $\sigma$, and sample size *N*) that constrain the means of samples from a population. The problem is the same for both conditions. The ideas that students in the discussion intuitively come up with are elaborated and formalised in subsequent problems.

**Problem 1. Expecting but not pregnant.**
Consider the following distributions of length:




- If someone was going to draw samples of 100 persons from the Dutch population, what do you think the **expected** value of the means would be?
- What would the **expected** value be of the means of samples from the populations of basketball players and elementary school children?

- Now suppose that we draw smaller samples (for example of only three persons) from these populations. What values for the means of these samples can be found?

Let's look at the fourth distribution of length.

- Suppose we draw samples of 100 persons from the population of Dutch soldiers. Which values of the means of these samples are less likely as compared to the means of the samples from the Dutch population?

*Can you indicate which 3 aspects determine your expectation of the mean of a sample that you have drawn?*

## 4.11 Appendix B

**Tutoring instruction of problem 1**
The tutors in both conditions were given some explanations about what the problem is supposed to evoke, what the core topics are, and some clarification of the subject matter.
In addition, the tutors in the guided condition received the following guiding questions.

**Example questions for the discussion (in the original frame the correct answers are also given).**

- *What is the problem about? Why does the problem have this title? Which question is asked several times?*
- *What are the values that can be expected? How high is the probability of a sample mean* ($\overline{X}$) *of 2.85? Can the means have any value? Look at the three populations.*
- *Do you expect exactly the population mean in a sample?*
- *When will* $\overline{X}$ *be higher than* μ *and when will it be smaller?*
- *Do you think that* $\overline{X}$ = 2.10 *is very likely from a sample of N = 100 from the Dutch population?*
- *Why is that unlikely?*
- *Could it be the* $\overline{X}$ *of a sample from the basketball players?*
- *Why is that?*
- *Give me ten values of people in a sample of N =10 from the population of Dutch soldiers. What do you expect for* $\overline{X}$ *?*
- *Now calculate* $\overline{X}$. *Is it exactly* μ*? Now substitute one of these ten values by 1.40, what happens to* $\overline{X}$ *?*
- *Was it possible to have a person of 1.40 in the sample?*
- *There are such small people in the Dutch population. What is your conclusion?*
- *Would it have had the same effect if we had substituted one of the values in a sample of n = 100? What is your conclusion?*
- *What can* $\overline{X}$ *be from a sample of the Dutch population if N =1; or N =3; or N = 10?*
- *Which three aspects determine which* $\overline{X}$*'s are likely?*
- *How are distributions of possible* $\overline{X}$*'s called?*
- *Which theorem constraints the possible* $\overline{X}$*'s ?*

## 4.12 Appendix C

### Evaluation questionnaire

Statements of the questionnaire with respect to respectively the course, the problems, the tutorial meetings, tutors stimulating understanding, and general functioning of the tutors.

*The course*

1. The goal of the course has become clear to me.
2. The course was instructive
3. The course was well organised

*The problems*

4. The problems showed sufficient variation
5. The problems evoked a systematic approach

*Elaboration*

6. The discussions in the meetings were effective
7. In the discussions the subject matter was integrated
8. The meetings were pleasant
9. The meetings were productive
10. The problems evoked useful learning goals
11. The problems evoked useful discussions

*The tutor stimulating understanding*

12. The tutor directed the discussions systematically
13. The tutor showed insight in the content of the course
14. The tutor helped structuring the subject matter
15. The tutors used their expertise to guide the group
16. The contributions of the tutors were relevant

*Aspects of the traditional role of the tutor*

17. The tutor stimulated autonomy
18. The tutor activated the students
19. The tutor regularly evaluated group processes

*Note*. These statements were rated by students on a five point Likert-scale

# Chapter 5

## The effect of distributed practice and directive tutor guidance on students' conceptual understanding of statistics

This chapter is submitted for publication as: The effect of distributed practice and directive tutor guidance on students' conceptual understanding of statistics, Luc Budé, Tjaart Imbos, Margaretha W.J. van de Wiel, Martijn P.F. Berger

## 5.1 Abstract

*Conceptual understanding depends on coherent, stable, and error free knowledge structures. The construction of such knowledge structures should be realised in education. The effects of reforming statistics education on students' conceptual understanding of the subject matter were examined in three consecutive studies. Curriculum changes in a naturalistic problem based learning environment enabled these studies, in which students were randomly sampled for measurement. To gauge students' understanding, they answered open ended questions in which was asked to explain and relate important statistical concepts. In the first study it was shown that distributed practice had a positive effect on students' understanding. In the second study it was shown that providing directive tutor guidance during tutor meetings improved understanding. Results of the third study showed that long term retention was limited.*

## 5.2 Introduction

In this paper the conceptual understanding of topics taught in statistics courses is investigated. A number of characteristics of conceptual understanding are reported in the literature. Conceptual understanding enables the explanation of causal mechanisms and processes (Graesser, Olde & Lu, 2001; Noordman & Vonk, 1998). It leads to better performance (Mayer, 1989) and better application and transfer of what has been learned (Novak, 2002; Feltovich, Spiro & Coulson, 1993). Conceptual understanding creates a sensation of coherence of the subject matter (Entwistle, 1995). This implies that stored knowledge elements are not isolated or arbitrarily connected. Conceptual understanding of studied material is only possible if knowledge elements are stored in coherent structures that contain all the relevant concepts of the domain, as well as their relations (Chi, Feltovich, & Glaser, 1981; Wyman & Randel, 1998; Kintsch, 1998). Our definition of an individuals' understanding is based on these findings. Conceptual understanding is shown, when a person demonstrates coherent, stable, error free knowledge structures. In this view conceptual understanding is related to the quality of the knowledge structures of an individual learner.

For example, the appropriate choice of statistical techniques in a study and correct interpretation of the results requires understanding of the material. In the case of a *t*-test, this means that one has to know why this should not be done on categorical data, whether the data are independent or not, how sample size is related to the power of the test, how a *p*-value should be interpreted in the light of a possible type I error, how the *p*-value and effect size are related to the relevancy of the findings, and what hypothesis can be tested. If an individual demonstrates correct knowledge of how these matters are interrelated, then we assume that person has conceptual understanding of the subject matter. Only being able to reproduce a definition, or literally quote a passage from a textbook, or mastering a skill without knowing why and when it is applicable, does not demonstrate conceptual understanding.

Conceptual understanding can be measured through the use of questions that require the combinations of concepts and questions that ask for the application of knowledge (Olson & Biolsi, 1991; McNamara, Kintsch, Butler-

Songer, & Kintsch, 1996). Open ended questions that ask for explanations, and underlying causal mechanisms are well suited for this purpose (Budé, van de Wiel, Imbos, Schmidt, Berger, 2006; Feltovich, et al., 1993; Gijbels, Dochy, van den Bossche & Segers, 2005; Jonassen, Beissner & Yacci, 1993). Answers to these questions can reveal how an individual links together the concepts. This discloses the knowledge structures of an individual and thus enables the assessment of conceptual understanding as we have defined it.

The development of conceptual understanding, i.e. the construction of coherent, stable, error free knowledge structures, is influenced by several factors. In three consecutive studies we will focus on two factors. The first factor, the effect of distributed practice on students' conceptual understanding, was investigated in a first study. In this study two conditions were compared. In one condition students had the opportunity to engage in distributed practice in a course that lasted for six months. In the other condition, because of a curriculum change, the students studied the same subject matter in a course with a time span of only eight weeks with considerable less opportunity for distributed practice.

The second factor that we investigated is whether providing additional directive guidance would enhance students' conceptual understanding. In a second study the effect of directive tutor guidance on conceptual understanding was investigated by comparing conditions with and without extra tutor guidance.

Finally, in a third study we investigated what the long term effect was of distributed practice and directive tutor guidance on how much students still understood of the subject matter, six months after the course.

*5.2.1 First study: Distributed practice versus massed practice*

A change in the curriculum programming in the academic year 2002-2003 at the faculty of Health Sciences of Maastricht University enabled us to compare two different statistics courses. Before the curriculum change, statistics was taught in courses with a time span of six months. The statistics courses included lectures, problem based learning (PBL) group meetings with a tutor in which subject matter was discussed, and practical sessions to train the students in using statistical software. Between these educational activities the students had at least one week of self-study. The intervals of one week or more between educational activities and the possibility to spread self-study over several days created more so-called distributed practice. Distributed practice is defined as study activities with intervals (e.g., Bahrick & Hall, 2005).

After the modification of the curriculum the statistics courses were concentrated in periods of eight weeks. As before the change, these condensed courses consisted of lectures, tutorial group meetings where the subject matter was discussed, and practical sessions. The lectures and practical sessions before and after the curriculum change were identical, the same books and literature were used, the PBL group meetings were similar, the number of educational activities was equal, students were instructed to study themselves the same material, and the courses covered the same topics. The introductory courses, in which this study was done, covered the central limit theorem,

sampling distributions, *t*-tests, analysis of variance (ANOVA), linear regression analysis, and methodological subjects such as randomised clinical trials and quasi experimental designs. So, in general it can be concluded that the courses before and after the change were comparable regarding content and implementation, and the total amount of time the students were supposed to spend on studying the subject matter.

The most important difference was that after the change the courses were much shorter and contained at least three educational activities each week. This reduced considerably both the spacing of educational activities and the opportunity to spread self-study. The students in the introductory courses had to master the above mentioned topics in the limited time span of eight weeks. Both the shorter intervals between educational activities and the reduced distribution of self-study led to more massed practice. Massed practice is defined as studying subject matter uninterruptedly or with only short breaks, in a brief time interval (e.g., Bahrick & Hall, 2005).

Beneficial effects of distributed practice in comparison to massed practice, i.e. the spacing effects, were found in research that at first predominantly was done in the laboratory and usually was focussed at memorising lists of items with intervals in terms of minutes or hours (Glenberg, 1979; Melton, 1970). In subsequent work the positive effect of distributed practice was also established in real educational settings (Seabrook, Brown, & Solity, 2005) and much longer time intervals (Bahrick & Hall, 2005; Bahrick & Phelps, 1987). Bahrick and Hall explain the spacing effect with students' metacognitive monitoring. That is, when study activities are spaced in time, students will notice that they have forgotten some of the material in the intervening period. When they experience such retrieval failures, they will be inclined to use encoding methods that will lead to better retention. They may also select and plan study strategies that will lead to less forgetting (Benjamin & Bird, 2006).

Based on the research on the distribution of practice it was expected that the massed practice that resulted from the curriculum change, would have a detrimental effect on the development of students' conceptual understanding of the subject matter. The research question of the first study was: will the reduction of distributed practice lead to less conceptual understanding of the subject matter that was taught in the introductory courses.

### *5.2.2 Second study: Directive tutor guidance*

In the second study another reform of the courses, with possible consequences for the conceptual understanding of the students, was investigated. This study focussed on the effect of an intervention toward the PBL tutorial meetings. The role of the tutor in the tutorial meetings is to stimulate the group processes, try to create an atmosphere in which students can optimally participate in the discussions, and to stimulate self-study (Barrows, 1988; Schmidt & Moust, 1995, 2000).

Expert tutors also tend to make subject matter contributions to the content of the discussion, which has been shown to improve novice student's performance (Schmidt, v.d. Arend, Moust, Kokx, Boon, 1993, Dolmans, et al.,

2002). Research has also shown that a tutor's expertise can compensate for lack of prior knowledge of students and that tutorial groups with relatively low levels of productivity require more input from a tutor (Dolmans & Wolfhagen, 2005). To be able to make subject matter contributions, the tutors usually have tutoring instructions at their disposal. These instructions consist of general information about the course, the main topics, more specific information about the goal of the problems, which subject matter is supposed to be discussed and how to tackle the problems. In our intervention the tendency of tutors to intervene in the content of the discussion, was taken a step further.

The intervention existed of additional tutor guidance during the meetings. That is, the tutors were given instructions to more actively guide the discussions in a directive way. The instructions consisted of written questions that the tutors could ask when appropriate, during the meetings. By asking these questions, the tutors could direct the discussion in the correct way as soon as the students would go astray or omit important subject matters.

The questions were based on how experts would deal with the presented problems. Step by step, in a presubscribed order the questions guided the students through the subject matter. The guiding questions focussed the discussions on relevant issues, indicating the direction of reasoning without hampering the active learning processes. Tutors received written directions and training in how to use the instructions before the start of the course. During the course each week the tutors discussed their experiences with this approach and they were given feedback and additional instructions in a meeting.

The intervention was assumed to enhance students' conceptual understanding of the subject matter, because students were provoked to focus on and actively think about the subject matter. Several studies have shown these positive effects of asking questions (Chi, Siler, Jeong, Yamauchi, & Hausman, 2001; Glaser, 1991; Graesser, Bagget, & Williams, 1996; Graesser, Person, & Magliano, 1995).

Additionally, tutoring as it was applied in the intervention is a way of coaching and modelling. It enabled reflection on and shaping of students' thinking and can therefore be considered as a form of process oriented learning (Vermunt, 1994; Volet, McGill, & Pears, 1995). Process oriented learning leads to better conceptual understanding (Vermunt & Vermetten, 2004; Volet, 1991). Moreover, by asking guiding questions the tutors may have focussed the employment of cognitive resources, thus reducing cognitive load. The reduction of cognitive load has been shown to positively affect learning of complex materials with a high level of element interactivity such as statistics (van Merriënboer & Sweller, 2005; Tuovinen & Sweller, 1999). Finally, by asking questions the tutors structured and directed the discussions. Structuring and directing the discussions reduced floundering, which is said to interfere with learning (Merrill, Reiser, Merrill, & Landes, 1995).

The intervention was assumed to be helpful especially in statistics, because a lot of students find statistics difficult and abstract, and the interconnectivity of the concepts is high (Schau & Mattern, 1997) Moreover, Garfield (2003) found that students often develop misconceptions while studying

statistics. The research question of the second study was whether directive tutorial guidance would have a positive effect on students' conceptual understanding of statistics.

*5.2.3 Third study: Long term retention*

The third study focussed at long term retention in two different educational settings, i.e. before and after the change in the curriculum in the academic year 2002-2003. In the years 2001-2002 and 2004-2005 it was measured how much students still understood of the subject matter six months after the course.

These measurements allowed us to compare long term retention of the students who followed a six months' course with students who followed a more condensed course. The six months' course offered the possibility for distributed practice. Research has shown that distributed practice leads to better retention (Bahrick & Hall, 2005; Bahrick & Phelps, 1987; Seabrook, et al., 2005). In the condensed courses additional tutor guidance was provided. The additional directive guidance could have, as explained above, caused better conceptual understanding (Glaser, 1991; Graesser, et al., 1996; Vermunt & Vermetten, 2004). Better conceptual understanding, based on more stable coherent knowledge structures, may cause an increased retention (Neisser, 1984; Semb & Ellis, 1994). The research question of the third study was how much conceptual understanding the students still had, six months after the two different courses.

*5.2.4 Design of the studies*

Data for the three studies were collected in independent samples of students from the population cohorts of 2001-2002, 2002-2003, and 2004-2005. All students were tested with the same open ended questions to measure conceptual understanding. This enabled a straightforward comparison within and between the cohorts. The questions were directed at statistical hypothesis testing theory.

Measurements were done at four different time points: before the start of the course, during the course, directly after the course, and six months after the course. The measurement during the course was immediately after the subject hypothesis testing was dealt with in the courses. In Table 5.1 the design for these three studies is presented.

It should be noted that recall and memory effects were controlled by independent sampling of students at different time points. In addition, members of the Department of Methodology and Statistics of Maastricht University were included in the study as an expert control group.

The paper is organised as follows: In the next paragraph study 1 is presented, with its results and discussion. Then, study 2 and study 3 are presented, respectively. Finally, overall conclusions are drawn.

Table 5.1. Design of the three studies with the cohorts from which the independent samples were drawn.

| *Populations from which was sampled* | *Study 1: Distributed vs massed practice* | *Study 2: Directive tutorial guidance* | *Study 3: Long term retention* |
|---|---|---|---|
| Cohort 1 | | | |
| 2001-2002 during the course | X | | |
| 2001-2002 after the course | X | | X |
| 2001-2002 after six months | | | X |
| Cohort 2 | | | |
| 2002-2003 during the course | X | X | |
| 2002-2003 during the course with extra guidance | | X | |
| 2002-2003 after the course | X | X | |
| 2002-2003 after the course with extra guidance | | X | |
| Cohort 3 | | | |
| 2004-2005 before the course | | | X |
| 2004-2005 after the course with extra guidance | | | X |
| 2004-2005 after six months with extra guidance | | | X |
| Members of the Dep. of Meth. & Statistics | X | X | X |

Notes: X indicates measurements. All measurements were done with the same open ended questions.

### 5.3 Study 1: The effect of distributed versus massed practice on conceptual understanding

The overall hypothesis in this study was that massed practice would have a detrimental effect on students' conceptual understanding of statistics. This is concretised into three hypotheses:

1. Distributed practice will lead to better conceptual understanding compared to massed practice.
2. Conceptual understanding will in both conditions be better after the course than during the course. Students are expected to refine their knowledge structures during the course, because the subject matter is repeatedly dealt with in the course. Therefore, it is expected that in both the distributed practice condition as in the massed practice condition the knowledge structures will be more stable, coherent, and error free at the end of the course.

3. The difference in conceptual understanding of the students in the distributed practice condition and the massed practice condition will be larger after the courses than during the courses. The effect of distributed practice will be the larger at the end of the course, because students will have had the largest amount of practice then. Thus, there will be an interaction between the kind of practice and the time factor.

*5.3.1 Participants*

Participants were first year bachelor students of the faculty of Health Sciences of Maastricht University, who were recruited during lectures and other educational activities. During recruitment they were told that they had to answer questions about statistics and that they would be paid 10 euro. This payment was given to avoid attracting only motivated students who were particularly interested in statistics. All students took the introductory statistics courses in which this study was executed. We only sampled students who volunteered to participate.

The students were sampled from two successive cohorts (the years 2001-2002 and 2002-2003) and randomly assigned to be measured at two measurement time points: during and directly after the course. One hundred and twenty two ($N$ = 122) students participated; sixty four in the year 2001-2002 ($N_{2002}$ = 64), and fifty eight in the year 2002-2003 ($N_{2003}$ = 58). Eighty-seven of these participants were female, 35 were male. Approximately 75 percent of the Health Sciences' students is female. The age of the participants ranged from 20 to 26 years. A separate group of nine faculty members of the Department of Methodology and Statistics ($N_{M\&S}$ = 9) participated as expert control group. Their age ranged from 25 to 60 years and their teaching experience ranged from 4 to 30 years.

*5.3.2 Measurement instrument*

To assess what the students had understood of the subject matter, ten open ended questions concerning statistical hypothesis testing were designed. These questions specifically asked for the explanations of and the relations between statistical concepts. For example, it was asked to explain the relation between the central limit theorem and the confidence interval. For a complete list of the questions see appendix A. A detailed and elaborated answer key with thesaurus was formulated with the aid of four statisticians. The answers were split up in propositions and awarded with a weighted scoring system. For reliability purposes a second rater scored a random subset of the data ($N$ = 14). Interrater agreement was high. The correlation between the scores was significant ($r$ = .950), the mean of the two raters did not significantly differ ($M_1$ = 20.07; $M_2$ = 18.36; $t$ = 1.20, $p$ = .250), nor did the standard deviation ($SD_1$ = 15.6; $SD_2$ = 16.7).

The total number of points that could be obtained was 110. This number could only be obtained if all the details of the quite comprehensive answer key were mentioned in the answers. An upper score limit was determined by the answers of a sample of nine faculty members of the Department of Methodology

and Statistics, who were considered to master the subject matter competently. Their average score served as a reference point.

In addition, students' final course exam grades were used. The grades served as an additional measure to test for the difference between the two courses. The exams of both the years 2001-2002 and 2002-2003 consisted of 30 multiple choice questions. Both exams measured students' achievement toward the subject matter.

*5.3.3 Procedure*

Students received written and oral instructions to answer the questions as completely as possible, even if they doubted whether an answer was correct. This was done to get information on partial knowledge and possible misconceptions. Answers had to be written on a blank sheet of paper. Time was not limited; the whole procedure took on average one hour.

*5.3.4 Design*

Independent samples of students, drawn from volunteer pools from both the cohorts 2001-2002 and 2002-2003, were measured during and directly after the course (see Table 5.1). In addition, the grades on the final course exam of the whole cohorts of 2001-2002 and 2002-2003 were registered. The sample of department members was measured in the year 2001-2002.

*5.3.5 Analysis*

The written answers were copied for scoring purposes. The answers to the questions were awarded with points in accordance with the scoring system. A two (*year*) x two (*time point*) fixed factor Analysis of Variance (ANOVA) was done with the score of the open ended questions as dependent variable. The grades on the final course exam of the two cohorts of 2001-2002 and 2002-2003 were compared with a *t*-test.

## 5.4 Results

The two (*year*) x two (*time point*) fixed factor Analysis of Variance (ANOVA) showed no significant interaction ($F(1,118) = .153$, $p < .697$). The model without the interaction term showed significant differences of both year ($F(1,119) = 36.05$, $p < .000$) and time point ($F(1,119) = 19.07$, $p < .000$). The mean of the faculty members differed significantly from all four groups ($p < .000$). The descriptive statistics of the measurements are presented in Table 5.2.

The grades on the final course exams of the entire cohorts of 2001-2002 and 2002-2003 were compared with a *t*-test. The results also showed a significant difference between both years ($M_{2002} = 6.45$, $M_{2003} = 5.9$, $t = 3.34$, $p < .001$).

Table 5.2. Descriptive statistics of the measurements of study 1.

| | *M* | *SD* | *Min* | *Max* | *N* |
|---|---|---|---|---|---|
| 2001-2002 during the course | 26.5 | 10.9 | 7.0 | 54.0 | 33 |
| 2001-2002 after the course | 35.3 | 10.4 | 12.0 | 61.0 | 31 |
| 2002-2003 during the course | 16.1 | 7.9 | 1.0 | 29.0 | 30 |
| 2002-2003 after the course | 23.4 | 11.2 | 4.0 | 46.0 | 28 |
| Faculty members | 51.2 | 19.2 | 17.0 | 73.0 | 9 |

Note: Measurements are independent, from random samples of the cohorts of students

## 5.5 Discussion

The first hypothesis, i.e. distributed practice will lead to better conceptual understanding, was confirmed. We found that the students measured during the course of 2001-2002 scored significantly higher than the students measured during the course of 2002-2003 and even better than the students measured in 2002-2003 after the course. The students measured in 2001-2002 after the course scored best. The change in the curriculum in the academic year 2002-2003 resulted in a considerable reduction of the time span of the statistics courses. The reduction of the distribution of the subject matter over time decreased the possibility of effectively using distributed practice. The results show that the loss of the spacing effect negatively affected students' conceptual understanding. This conclusion is supported by the grades on the final course exams. Both exams were supposed to assess the same level of understanding of the subject matter. The mean of the whole cohort of 2001-2002 altogether was significantly higher than the mean of the cohort of 2002-2003.

With regard to the second hypothesis, it was found that in both the cohorts of students in the years 2001-2002 and 2002-2003, conceptual understanding improved significantly during the course. This means that the knowledge structures of the students gradually became more stable, coherent, and error free during the course.

The third hypothesis, i.e. the difference in conceptual understanding between the students in the distributed practice course and the massed practice course will be larger after the courses than during the courses, was not confirmed. Despite the fact that the interaction was not significant, it was in the expected direction. The improvement of conceptual understanding over the course was somewhat less in 2002-2003 than in 2001-2002.

The scores of the faculty members were taken as an upper bound. Faculty members answered the questions unprepared. In contrast, the students answered the questions after they had studied the subject matter. However, at the end of the introductory courses students in both conditions, as expected, clearly performed below the level of the faculty members.

Although our results, indicating a positive effect of distributed practice, corroborate previous research (Bahrick & Hall, 2005; Bahrick & Phelps, 1987;

Seabrook, et al., 2005), there may be a threat to the internal validity of this quasi-experimental study. Even though there was no reason to assume that the cohorts of students of 2001-2002 and 2002-2003 differed with respect to their potential capacity to master the subject matter, or with respect to their prior knowledge, this is a possibility that can not be completely ruled out. We did, however, compare the cohorts of 2002-2003, 2003-2004, and 2004-2005 on their overall pass rates with respect to the other courses of the first year and we did not find significant differences. This indicated that these cohorts were comparable. The pass rate data for the cohort of 2001-2002 were not available.

## 5.6 Study 2: The effect of directive tutor guidance on conceptual understanding

Based on the literature reviewed, the overall research hypothesis of this study is: directive guidance will have a positive effect on students' conceptual understanding of the subject matter. More specifically, we have three hypotheses:

1. The intervention group, with directive tutor guidance, will have a better conceptual understanding than the control group.
2. Conceptual understanding will be better after the course than during the course, as knowledge structures will gradually develop during the course.
3. At the end of the course the effect of directive guidance will be maximal. Therefore, the differential effect of directive tutoring will be larger after the course, than during the course, i.e., we assume an interaction effect between tutor guidance and time points.

### *5.6.1 Participants*

The whole cohort of 2002-2003 of first year Bachelor students of the Faculty of Health Sciences at the University of Maastricht was involved in this study. The cohort, enrolled in a PBL introductory statistics course, consisted of 208 students. Before the start of the course 138 volunteers were recruited from this cohort in the same way as in Study 1; 66 in the guided condition, 72 volunteers in the control condition. One hundred and eleven students were randomly sampled from these volunteers and assigned to two measurement time points, during and after the course. These 111 students were measured with respect to their conceptual understanding.

### *5.6.2 Measurement instrument, procedure, design, and analysis*

All 208 students were randomly assigned to 24 tutorial groups. These 24 tutorial groups were randomly assigned to 14 tutors (ten tutors had two groups, four had one group). Finally, the tutors were at random assigned to either the guided condition or to the control condition using blocked randomisation. This resulted in 12 tutor groups ($N$=103) being assigned to the guided condition and 12 tutor groups ($N$=105) to the control condition. The tutors were members from the Department of Epidemiology and the Department of Methodology and Statistics. All tutors had sufficient knowledge of the subject matter.

The same questions with the same scoring system and procedure were used as in study 1. These questions were used to measure conceptual understanding of those students who volunteered. Students of both the conditions (*condition*) were measured at the same time points (*time points*): during and directly after the course. A two-way analysis of variance (ANOVA) was done to compare the means. As in study 1 the score of the faculty members served as a reference point. In addition, students' final course exam grades were used. The grades served as an additional measure to test for the difference between the two conditions. The mean grades of the students in the two conditions were compared with a t-test. The exam consisted of 30 multiple choice questions, measuring students' achievement toward the subject matter.

### 5.7 Results

The two (*condition*) x two (*time point*) fixed factor ANOVA showed no significant interaction ($F(1,107) = .309$, $p < .579$). The model without the interaction term showed significant main effects for *condition* ($F(1,108) = 18.681$, $p < .000$) and *time point* ($F(1,108) = 12.21$, $p = .001$). Again the mean of the faculty members differed significantly from all other groups ($p < .000$). The descriptive statistics of the measurements are presented in Table 5.3.

Table 5.3. Descriptive statistics of the measurements of study 2.

| | *M* | *SD* | *Min* | *Max* | *N* |
|---|---|---|---|---|---|
| During the course control condition | 16.13 | 7.93 | 1.0 | 29.0 | 30 |
| After the course control condition | 23.43 | 11.23 | 4.0 | 46.0 | 28 |
| During the course guided condition | 24.93 | 10.04 | 8.0 | 49.0 | 28 |
| After the course guided condition | 30.20 | 8.75 | 15.0 | 46.0 | 25 |
| Faculty members | 51.22 | 19.27 | 17.0 | 73.0 | 9 |

Note: Measurements are independent, from random samples of the cohort of students in the year 2002-2003

The grades on the final course exam of both conditions were compared with a *t*-test. The results of the *t*-test showed a marginally significant difference between both conditions ($M_{control} = 5.85$, $M_{guided} = 6.20$, $t = 1.47$, $p < .072$).

### 5.8 Discussion

The first hypothesis, i.e. the intervention group will have a better conceptual understanding than the control group, was confirmed. We found that in the guided condition the students scored considerably better than in the control condition. It is remarkable that the students in the control condition never reached the level of the students in the guided condition. Providing directive guidance seems to be an effective addition to problem based learning of statistics. This finding is supported by the grades on the final course exam. The

students in the guided condition scored better than the students in the control condition.

The second hypothesis, i.e. conceptual understanding will improve over the course, was also confirmed. This is in line with the results of study 1. However, the gradual development of conceptual understanding was slower in the control condition. Students in the control condition scored poorly during the course ($M$ = 16.13), showing a low level of conceptual understanding, whereas students in the guided condition scored already rather well ($M$ = 24.93). Directive tutor guidance not only seemed to have improved the final level of conceptual understanding, but also seemed to have promoted a more gradual development of conceptual understanding.

The difference between the conditions was smaller directly after the course than during the course. This interaction was not significant, and in the opposite direction as hypothesised (hypothesis 3). Toward the end of the courses students usually study hard in preparation for the final course exam. Apparently, the students in the control condition studied harder in the last week of the course, but could only partially catch up.

Compared to the reference point, i.e. the score of the faculty members, all of the groups, as expected, scored far below the level of the faculty members. As in study 1 this obviously shows that before reaching a solid level of conceptual understanding of statistics, students have to spend more time studying the subject matter and need more experience.

## 5.9 Study 3: Long term retention

The overall research hypothesis of this study is: distributed practice and directive tutor guidance will have a similar effect on what students still understand of the subject matter, half a year after the end of the course. Both factors have shown to enhance conceptual understanding during and directly after the course (see study 1 and 2). We assume that these factors will have the same effect on conceptual understanding six months after the course too. More specifically we have four hypotheses:

1. Students in the directive tutor guidance condition and in the distributed practice condition will have the same level of conceptual understanding directly after the course and after six months.
2. Conceptual understanding will be better directly after the courses than after six months, as students will probably forget some of the subject matter.
3. Conceptual understanding of both conditions will be better six months after the courses than before the start of the courses.
4. There will be no differential effect for both conditions; we do not expect an interaction between the conditions and the time points.

### *5.9.1 Participants*

From the pool of volunteers, which was obtained in study 1, 30 students were at random allocated to the measurement directly after the course, 31 to the long term measurement of 2001-2002. Before the start of the course 2004-2005, 75 volunteers were recruited from the first year students in the same way as in

study 1. From these volunteers, 23 students were at random selected for the measurement of the entry level of understanding, 31 students directly after the course, and 24 students for the long term measurement. For practical reasons it was not possible to measure long term retention in the intermediate years.

*5.9.2 Measurement instrument, procedure, design and analysis*

The same measurement instrument was used as in studies 1 and 2. This enabled a comparison within a specific year, as well as a comparison between years. Moreover, the same procedure was followed as in studies 1 and 2. Random samples of students of both the cohorts in 2001-2001 and 2004-2005 (*year*) were measured directly after the course and six months after the conclusion of the course (*time point*). A two-way analysis of variance (ANOVA) was done to compare the means. For practical reasons it was not possible to measure the entry level, i.e. prior knowledge at the beginning of the course, in the year 2001-2002. In 2004-2005, however, a third independent sample of students was used to measure the entry level. The scores of the long term measurements were also compared with this entry level score with independent samples *t*-tests.

## 5.10 Results

The two (*year*) x two (*time point*) fixed factor ANOVA showed no significant interaction ($F(1,112) = 1.947$, $p < .166$). The model without the interaction term showed a significant difference of time point ($F(1,113) = 88.90$, $p < .000$), but not of year ($F(1,113) = .73$, $p = .393$). Compared with the entry level score both long term measurements were significantly higher ( $p_{2001\text{-}2002} < .000$, $p_{2004\text{-}2005} < .000$). The respective means are presented in Table 5.4.

Table 5.4. Descriptive statistics of the measurements of study 3.

| | *M* | *SD* | *Min* | *Max* | *N* |
|---|---|---|---|---|---|
| 2001-2002 after the course | 35.26 | 10.42 | 12.0 | 61.0 | 31 |
| 2001-2002 after six months | 14.73 | 7.74 | 3.0 | 27.0 | 30 |
| 2004-2005 after the course | 31.16 | 12.54 | 5.0 | 57.0 | 31 |
| 2004-2005 after six moths | 15.96 | 9.25 | 4.0 | 33.0 | 24 |
| 2004-2005 before the course | 6.13 | 4.48 | 1.0 | 18.0 | 23 |

Note: Measurements are independent, from random samples of the cohorts of students

## 5.11 Discussion

The first hypothesis was that students in the distributed practice condition and in the directive tutor guidance condition would have the same level of conceptual understanding at the two measured time points. This hypothesis was confirmed by our results; conceptual understanding of students of both cohorts is comparable when measured directly after the course as well as after six months.

In the year 2001-2002 the course was spread out over a semester. This created a natural spacing effect, which had a positive effect on conceptual understanding (see study 1). In the year 2004-2005 the tutors provided extra guidance to the students, which had shown to have a beneficial effect on students' understanding in the year 2002-2003 (see study 2).

The comparison of long term retention between the year 2001-2002 and year 2004-2005 showed no significant difference. The loss of the natural spacing effect of the courses after the change in curriculum in the academic year 2002/2003, seems to have been neutralised by the directive guidance that the tutors provided in the courses of subsequent years.

The second hypothesis was also confirmed. There was a significant decline in the level of conceptual understanding in the six months between the two measurement time points. Although this decline was expected and hypothesised, the actual magnitude of this fall back was not foreseen.

With regard to the third hypothesis, there was a significant improvement of conceptual understanding six months after the courses, when compared to the entry level of 2004-2005. However, this improvement was rather limited (see Table 5.4). Moreover, the level of students' conceptual understanding six months after the course was far below the level of understanding of the faculty members. The findings regarding the second and third hypothesis support the idea that students have to study a lot more and need more experience, before reaching a high level of conceptual understanding.

The results showed no significant interaction between the cohorts and the time points, as expected (hypothesis 4).

### 5.12 Final remarks and conclusions

Several conclusions can be drawn from these studies. Firstly, distributed practice enhances conceptual understanding of statistics. Secondly, directive tutor guidance has a positive effect on students' conceptual understanding of statistics. Asking the right questions at the right time seems to have activated the students, focussed them on the important topics, and stimulated mental activities especially at points where they had difficulties. This approach enhanced the level of students' conceptual understanding. Our result is in line with the idea of Graesser, et al. (1995), that the vision of learners as active, self-motivated, and inquisitive individuals is a bit too optimistic. Only providing the facilities for optimal learning is not enough (Vermetten, Vermunt, & Lodewijks, 2002). It seems necessary to activate students and tutor guidance seems to be a suitable approach in this respect.

On the other hand, the level of conceptual understanding of the students in all the cohorts was far below the expert level of understanding. It was not expected that after an introductory statistics course the students would be able to reach an expert level. In fact, this cannot be expected after any introductory course, regardless of the content. However, in all three studies the difference between the scores of the experts and the students was rather large. Moreover, in our third study of long term retention it was shown that the level of conceptual understanding substantially declined as time passed. Students' average score six

months after the course is 15 to 20 points lower than directly after the course and only 9 to 10 points better than on the entry level test (see Table 5.4). The low long term performance of the students is problematic, considering that after the final course exam, in most cases the subject matter does not recur. Yet, in subsequent statistics courses the subject matter of previous courses is assumed to be known. Without this prior knowledge it will be impossible for students to advance to higher levels of conceptual understanding. It seems to be advisable, therefore, to try to improve students' conceptual understanding even after the statistics courses are finished. Distributed practice might be a fruitful approach in this respect. It could be demanded from students to apply in subsequent non-statistical courses what they have learned in the statistics courses. For example, when discussing scientific literature, the statistical analyses of the studies and the results could be scrutinised. This would cause the subject matter of the statistics courses to recur regularly. It has been shown that without a periodic recurrence of the subject matter, the relations between concepts will fade and conceptual understanding will decline (Conway, Cohen, & Stanhope, 1991, 1992; Semb & Ellis, 1994). Future research is needed to study if and how this can be implemented and whether such an approach would lead to the anticipated positive effect on long term retention. The methodology most appropriate for such research might be design-based research, in which systematically adjusted aspects of education can be tested in naturalistic contexts (Barab, & Squire, 2004).

The instrument for the measurement of conceptual understanding consisted of ten open ended questions, which were scored on the basis of a comprehensive answer key. This answer key contained all elements of the possible answers down to the smallest detail. It was not expected that it would be possible to answer the questions so completely that the maximum of 110 points could be obtained. Although time was not limited, the available time appeared to be insufficient to reproduce all minutiae. The answers of the faculty members therefore served as an upper limit. Scoring the answers to the open ended questions with the answer key, in combination with this reference point seemed to be an appropriate measurement of students' conceptual understanding.

All three quasi-experimental studies presented in this paper were executed in an educational field setting. This might limit the scope of the conclusions somewhat, because we could not control all threats to internal validity. For example, it could be questioned, whether students in the distributed practice condition actually have studied regularly during the course, because we did not measure self-study. We are confident, however, that students did practice in a distributed way, because attending the meetings, in which exercises were discussed, were obligatory for the students. Attending these meetings every two weeks and discussing the subject matter created the distributed practice.

Regarding the tutor guidance in the second study, we did not collect data during the tutorial meetings. This means that we do not have actual data on how the discussions developed, i.e. whether the tutors succeeded in directing the discussions by asking questions. However, we did discuss the experiences of the tutors with this approach every week in a meeting, in which feedback was given

as well. On account of the tutors' reports we are confident that the intervention was properly implemented.

In study 2 the effect of additional tutor guidance on conceptual understanding was investigated. As explained in the introduction, the positive effect of guidance was supposed to be mediated by students' elaborations on critical points, deep reasoning, explanatory reasoning, increased interactivity of tutoring dialogues, process oriented learning, increased social interaction, decreased cognitive load, and reduced floundering. However, none of these intervening processes and mechanisms were measured or studied. It would be interesting to investigate these mechanisms in future research. For example, it could be studied how the actual discussions in the group meetings take place and whether guiding questions effectively enhance reasoning processes and reduce cognitive load.

In conclusion, we are confident that our findings reflect genuine effects, because in all three studies the students were randomly sampled from the volunteer pools and assigned to the measurement time points. Moreover, as the studied conditions were embedded in realistic learning settings, it may be expected that the external validity of the outcomes is relatively high.

## 5.14 Appendix

### List of questions

1. **Explain** the relation between the central limit theorem and the confidence interval. First **explain** both terms and then how they are related.
2. Suppose that a 95% confidence interval for $\mu$ is 85 to 105. Can it be assumed then, with 95 % confidence, that $\mu$ can range between 85 and 105? **Explain** your answer.
3. **Explain** the formula: $\bar{x} \pm z_{1-\alpha/2}(\sigma/\sqrt{n})$. First **clarify** the different terms and then the formula (i.e. **explain** why the terms are in the formula).

Suppose a researcher studies the IQ of students. The mean IQ of the Dutch population is known to be 100. The hypotheses of the study are: $H_0$: μ = 100; $H_1$: μ > 100. A random sample is drawn from the population of students. The sample mean is 120.

4. **Explain** in your own words which procedure is needed in the example above to be able to reject or not reject the null hypothesis. **Explain** why.
5. Suppose the researcher, in the example above, decides to reject the null hypothesis. Is it certain then, that the alternative hypothesis is true? **Explain** your answer.
6. Suppose another researcher uses the same data as in the example above. This researcher does not reject the null hypothesis (without making a mistake). **Explain** what he can have done differently.
7. Suppose a researcher reports that the sample mean does not exactly match the hypothesised population mean. **Explain why** it is more informative to report the confidence interval and what extra information it provides.
8. **Explain** what a type II error is and when the probability of such an error increases.
9. **Explain** in your own words what a test statistic is and how it is used in tests.
10. When an absolute *z*-value increases, then the corresponding *p*-value decreases. **Explain** why, using the concept of normal distribution.

# Chapter 6

## The effect of guiding questions on students' performance and attitude toward statistics

This chapter is submitted for publication as: The effect of guiding questions on students' performance and attitude toward statistics Luc Budé, Margaretha W.J. van de Wiel, Tjaart Imbos, Martijn P.F. Berger.

## 6.1 Abstract

*In this study the effect of guidance on students' performance was investigated. This effect was hypothesised to be manifested through a reduction of cognitive load and enhancement of self-explanations. In an experimental setting two randomly selected groups of students answered achievement and transfer questions on statistics as a measure of performance. Students in the intervention condition were given guiding questions to direct their way of reasoning before they answered the achievement questions. The students in the control condition were only asked to write down their way of thinking before they answered the same achievement questions. The correct answers to the guiding questions and the reported correct steps in the way of thinking served as a measure of self-explanations. It was found that students in the intervention condition performed significantly better on achievement and transfer questions and that this effect of guidance was mediated by self-explanations and cognitive load. Attitude towards statistics was positively related to performance.*

## 6.2 Introduction

The purpose of this study was to investigate the effect of guidance on students' performance in statistics. In general, evidence from empirical studies has favoured a guided approach to learning over unguided instruction (Kirschner, Sweller, & Clark, 2006; Mayer, 2004). Guidance during instruction has proven to be effective with regard to students' performance both in research on human tutoring (Chi, 1996) and in research on worked examples (Tuovinen, & Sweller, 1999). In the educational context of problem based learning Budé, Imbos, van de Wiel, Broers, and Berger (2005) have shown that extra guidance provided by the tutor improved students' performance of statistics.

The positive effects of guidance may be attributed to a reduced cognitive load (van Merriënboer, & Sweller, 2005) and to enhanced self-explanations (Chi, 1996; Renkl, 2002).

## 6.3 Cognitive load

Cognitive load can be defined as the amount of effort needed by the human cognitive system to process information (Sweller, 1988). The capacity to process information is limited especially in the working memory (Baddeley, 1992, 2000). Cognitive load is high when subject matter makes high demands on the working memory. Instruction methods that take the limited capacity of the working memory into account are aimed at reducing cognitive load. As a result of this reduction the working memory can more optimally be employed for processes relevant to learning and for acquisition of schemata (Sweller, van Merrienboer, & Paas, 1998; van Merriënboer, & Sweller, 2005), thus leading to a better performance (e.g., van Gog, Paas, & van Merriënboer, 2006; Renkl, 2002; Tuovinen, & Sweller, 1999)

Carroll (1994) found that the reduction of cognitive load improved learning of mathematics. Sweller, et al. (1998) claim that the reduction of cognitive load also improved the performance in complex knowledge domains in which learning involves simultaneous processing of various interacting elements, such as

science, technology and computer programming. Paas (1992) found that students' transfer performance in statistics improved due to reduced cognitive load. Students with limited prior knowledge profit most from instruction methods that are aimed at reducing cognitive load (Kalyuga, Chandler, Tuovinen, & Sweller, 2001, Tuovinen, & Sweller, 1999).

On the basis of these findings it was expected that reducing cognitive load would lead to improvement of students' performance in statistics.

### 6.4 Self-explanations

Self-explanations can be defined as the internal active explanation of the learner of the steps in the reasoning process to answer a question or to solve a problem. It has been observed that the majority of the learners do not spontaneously engage themselves in self-explanations (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Renkl, 1999).

Stimulating or prompting self-explanations increases cognitive activity. Relevant cognitive activity promotes the construction of coherent knowledge structures, which enhances understanding of the subject matter and performance of the learner (Chi, 1996; Chi, Siler, Jeong, Yamauchi, & Hausman, 2001). Renkl and colleagues have shown that the best possible learning can be realised by combining the elicitation of self-explanations with the reduction of cognitive load (Renkl, 1997, 1999, 2002; Renkl, Stark, Gruber, & Mandl, 1998).

### 6.5 The model

In the present experimental study, the effect of guidance on performance in statistics via the reduction of cognitive load and stimulation of self-explanations was investigated. The model presented in Figure 6.1 shows the hypothesised relations between these constructs and their effects.

The model portrays that guidance is expected to reduce cognitive load and elicit more self-explanations. Both effects are hypothesised to enhance performance. High performance, in turn, is expected to be related to a more positive attitude toward statistics.

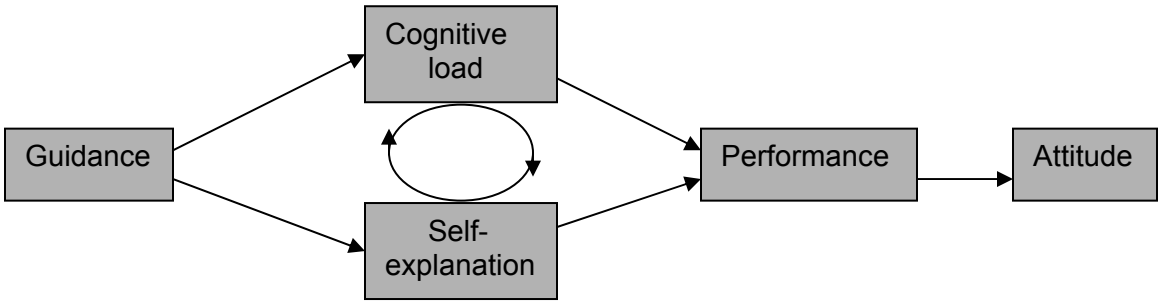



Figure 6.1. The effect of guidance on performance and attitude.

The relation between cognitive load and self-explanations can be characterised as recursive. The reduction of cognitive load enables the working memory to be used to a greater extent for self-explanations. The increased

number of self-explanations, i.e. increased cognitive activity, recursively heightens cognitive load. On the other hand, explaining to oneself the successive steps in a line of reasoning may activate more productive cognitive activity and may reduce floundering (Mayer, 2004). This in turn may reduce cognitive load (Sweller, 1988; van Merriënboer, & Sweller, 2005). The circular arrows in the model express this recursive relation.

### 6.6 Guidance

Research on reducing cognitive load in instruction mainly focussed on worked examples (e.g., Tuovinen & Sweller, 1999; Renkl, 2002; Gerjets, Scheiter, & Catrambone, 2006). In worked examples (some of) the successive steps that are needed to come to the solution of a problem are given. Worked examples can be considered a form of guidance, because they guide the learner from the initial problem through the solution steps to the final solution of a problem. Learners do not have to discover the solution steps themselves and this will reduce their cognitive load.

The guidance that was afforded to the students in the present study resembles worked examples. It consisted of providing directive questions which were aimed at successive reasoning steps. The line of reasoning of the students in the intervention condition was thus guided by the directive questions. In this way the questions were supposed to reduce floundering and, analogous to the effect of worked examples, to reduce cognitive load.

The present study differed from research on worked examples for two reasons. We did not give students problems to solve, but open ended questions. These questions offered the opportunity to ask the students to explain the relations between the concepts in their own words. We neither provided them with information about the correct answers, as is usually done in research on worked examples (e.g., Renkl, 1999, 2002; Ward, & Sweller, 1990). We did not want to spoon-feed the students with the correct answers, but we wanted the students to come up with the correct answers themselves and thus to actively apply and expand their knowledge. On the one hand this decreased the amount of guidance compared to worked examples, on the other hand it was expected that this approach would stimulate the students to engage in self-explanations to a greater degree.

Human tutors can also guide student learning by prompting self-explanations. By means of successive dialogue exchanges between the tutor and the students, and by focussing students' attention, the tutor can scaffold the construction of self-explanations, which in turn promotes learning (Chi, 1996; Chi, et al., 2001; Merrill, Reiser, Merrill, & Landes, 1995; Merrill, Reiser, Ranney, & Trafton, 1992). The intervention with the directive questions mimicked human tutors who guide student learning, because it focussed students' attention and was expected to elicit self-explanations.

Our approach is different from human tutoring, because no dialogue interaction between the students and their tutor took place. Dialogue interaction and individual differences of tutors are difficult to control in an experimental study

and can cause confounding effects. To eliminate this threat to internal validity we standardised guidance by using only written directive questions.

The directive questions were expected to combine the best of both worked examples and human tutoring, respectively the reduction of cognitive load and the promotion of self-explanations. However, by only providing directive questions instead of spoon-feeding the correct answers we diminished the amount of guidance. The standardised guidance through the directive questions was also less flexible than that of human tutoring. The question was whether this trimmed version of guidance would still reduce cognitive load, stimulate self-explanations, and enhance students' performance.

Finally, providing directive questions may cause the students to spend more time on the task. Time-on-task is also a possible confounder in research on student learning (e.g., Chi, et al., 1989; Renkl, 1997). The time that the students needed was therefore registered in the present study.

### 6.7 Performance

Performance was measured in two ways, with achievement and transfer questions. Both were open ended questions intended to measure higher levels of understanding of statistical concepts. The achievement questions required explaining the relations between different statistical techniques in a hypothetical study and interpreting the results of the statistical analyses. Answering questions that require the explanation of the relations between several concepts and the explanation of the applicability of those concepts is only possible when students have reached a high level of understanding (Dochy, 2001; Feltovich, Spiro & Coulson, 1993; Gijbels, Dochy, van den Bossche & Segers, 2005; Jonassen, Beissner & Yacci, 1993).

In addition to these achievement questions, performance was also measured with transfer questions, i.e. questions covering similar but slightly differently presented statistical topics. Transfer questions can be regarded as the ultimate measure of understanding. Previous studies have shown that performance on transfer questions is better when students developed a high level of understanding (Barnett & Ceci, 2002; Catrambone, 1998; Mayer, 1989; Olson & Biolsi, 1991).

### 6.8 Efficiency

Guidance is expected to reduce cognitive load and to enhance self-explanations. Enhanced self-explanations were expected to recursively raise cognitive load. The resulting level of cognitive load could therefore be higher or lower due to the guidance. However, the level of cognitive load per se is not very interesting. The combination with performance is important. Efficiency is a concept that combines performance with perceived cognitive load (Paas & van Merriënboer, 1993, Tuovinen, & Paas, 2004). When students in one group perceive the same cognitive load as students in another group, but their performance is higher, then their efficiency is higher. Alternatively, when students in one group perform equal to students in another group, but their cognitive load

is lower, then their efficiency is also higher. In the present study it was expected that due to the guidance the efficiency would be raised.

## 6.9 Attitude

Attitudes and beliefs of learners are rarely studied in research on the effect of guidance during instruction. However, attitudes and beliefs are known to affect students' learning of statistics (Gal, Ginsberg, & Schau, 1997). Negative events can trigger a downward spiral of an increasingly negative attitude (Graham & Weiner, 1987; Peterson, Maier, & Seligman, 1993; Pintrich & Schunk, 1996; Weiner, 1986, 1992). For example, Budé, et al. (2006) found that a negative attitude may impede learning of statistics. Unsuccessful learning and failing exams in turn will lead to an even more negative attitude. In the present study it was investigated whether a positive experience would lead to a more positive attitude. More specifically, the questions was whether guidance via an enhanced performance would lead to a more positive attitude toward statistics. This effect is displayed on the right hand side of the model in Figure 6.1.

In the present study the effects displayed by the model were investigated. The research hypotheses were:

- Guidance will improve performance via reduced cognitive load and enhanced self-explanations.
- Guidance will raise the efficiency of the students' learning.
- Guidance will via a higher performance lead to a more positive attitude toward statistics.

## 6.10 Method

### *6.10.1 Participants*

Forty nine (*N* = 49) second year bachelor students from the faculty of Health Sciences participated in this study, 39 of these participants were female, 10 were male. Approximately 75 percent of the Health Sciences' students is female. The age of the participants ranged from 20 to 26 years. The participants were recruited during educational activities. They were told that they had to fill in questions about statistics and that they would be paid 10 euro. This payment was given to avoid attracting only motivated students who are particularly interested in statistics. All participants had followed an introductory statistics course a half year before the study, i.e. they had previously studied the topics that were questioned in the present study.

### *6.10.2 Design and procedure*

Students were randomly assigned to two conditions. Before the start of the experiment, students in both conditions received written and oral instructions about how to answer the questions and fill in an attitude questionnaire. Students in both conditions then answered ten achievement questions. In the intervention condition guidance was afforded to the students by providing guiding questions prior to each achievement question. In the control condition the students were only asked to report their thinking steps before they answered the achievement questions. After each achievement question students rated the cognitive load

they perceived with regard to answering the question. That is, in the intervention condition the students first answered the guiding questions, then answered the achievement question, and finally rated the perceived cognitive load. In the control condition students analogously first reported their thinking steps and then answered the achievement question, followed by the cognitive load rating. After completing the achievement questions, students in both conditions answered the same transfer questions. Finally, they filled in a questionnaire measuring their attitude toward statistics. Answering the achievement questions took on average approximately 50 minutes, with minimum time 38 minutes and longest time 105 minutes. There was no time limit. Answering the transfer questions and filling in the attitude questionnaire took on average another 15 minutes. Students were tested in small group sessions ranging from four to eight people. The sessions were either of the intervention or the control condition. On completion of the attitude questionnaire they received the award of €10,-.

## 6.11 Materials

*6.11.1 Guiding questions.*

For the intervention condition directive questions were formulated. These directive questions helped the students to gradually build up their line of reasoning. They focussed the students on relevant issues, indicating the direction of reasoning. Students had to explain each step themselves; no extra information was given about the correct answers to the achievement questions. For examples of guiding questions see appendix A. Each correct answer to the guiding questions was regarded as one self-explanation. The students in the control condition were only asked to write down the steps of their line of reasoning in answering the achievement questions. For these students each correct step in their reasoning was regarded as one self-explanation.

*6.11.2 Achievement questions.*

The achievement questions related to a hypothetical health sciences study in which two independent groups were compared. The difference between the means of the two groups was analysed with three techniques, a *t*-test, a regression analysis and a one way analysis of variance. The hypothetical study and the results of the analyses were presented to the students. Ten open ended achievement questions regarding this study were formulated. They asked for analogies and differences of the three techniques and the interpretation of the results of the analyses. Correct answers required relating the three statistical techniques to each other, to the study, or to the analysis results.

The hypothetical study and examples of the questions are also given in appendix A. Students in both conditions received the same achievement questions. Their answers were scored blindly making use of an answer key.

*6.11.3 Transfer questions.*

Three open ended transfer questions were formulated. They related to a similar hypothetical study and statistical techniques for the comparison of two independent groups. However, a different research question, different figures

and different *p*-values were presented. For this study and an example question see appendix B. The transfer questions were identical for both conditions. The answers were also scored blindly making use of an answer key.

*6.11.4 Mental effort rating.*

Cognitive load was operationalised as the intensity of mental effort being expended by students. Students rated this mental effort on a 9-point Likert scale, responding to the following question: *Indicate how much mental effort it demanded to answer the above question*. The labels assigned to the numerical values of the scale ranged from: very, very low mental effort (1), neither low nor high mental effort (5), to very, very high mental effort (9).

This scale for mental effort can be used as an indicator of cognitive load (Paas, 1992, Paas, & van Merriënboer, 1993), and has shown to be a reliable and valid measure (Gimino, 2002; Paas, Tuovinen, Tabbers, & Van Gerven, 2003; Paas, van Merriënboer, & Adam, 1994). After each achievement question it was asked how much mental effort the answer to that particular question required.

*6.11.5 Attitude questionnaire.*

A modified Survey of Attitudes Toward Statistics (SATS; Gal, et al., 1997) was used to measure the attitudes toward statistics. It consisted of statements that students had to rate on a 7-point Likert scale. The original statements focussed on studying statistics in general and the adapted statements focussed on the participation of the students in the present study.

The questionnaire, like the SATS, comprises four subscales, measuring separate aspects of the attitude toward statistics. These four aspects are: value, difficulty, affect, and cognitive competence. Statements regarding value are focussed at usefulness, relevance and worth of statistics. Statements regarding difficulty focus at the difficulty of statistics. Statements regarding affect were directed at emotional aspects and feelings toward statistics. Finally, statements regarding cognitive competence focussed at students' intellectual ability and skills with respect to statistics.

For the statements of the attitude questionnaire, including the adapted statements, see appendix C.

## 6.12 Analysis

*Variables.* The independent variable *Guidance* consists of two levels: the control condition and the intervention condition. *Achievement Questions* consists of the sum scores of the ten achievement questions, ranging from 0 to 28. *Transfer Questions* consists of the sum scores of the three transfer questions, ranging from 0 to 12. *Mental Effort* is composed of the mental effort ratings. *Self-Explanations* comprises the correct answers to the guiding questions of the students in the intervention condition and the counted correct steps in the lines of reasoning of the students in the control condition. *Performance* stands for the overall performance score and was obtained by summing up *Achievement Questions* and *Transfer Questions*. The four subscales of the attitude

questionnaire constitute: *Value, Difficulty, Affect,* and *Cognitive Competence*. *Attitude* is the sum score of the four subscales.

*Efficiency* was calculated with the formula given by Tuovinen and Paas (2004), where $Z_{(\text{Achievement Questions})}$ and $Z_{(\text{Mental Effort})}$ are the standardised scores for these variables:

$$Efficiency = \frac{Z_{(Achievement\ Questions)} - Z_{(Mental\ Effort)}}{\sqrt{2}}$$

*Time* stands for the registered time needed for students in both conditions to answer the questions and to fill in the attitude questionnaire. Finally, *Grades* are the students' grades on a preceding statistics course exam.

*Statistical analyses.* The first hypothesis was that guidance via cognitive load and self-explanations would improve performance. To confirm this hypothesis, direct and indirect effects of *Guidance* were tested, i.e. once without and once with controlling for mediating variables. To test for the direct effects of *Guidance* on the *Achievement Questions*, *Transfer Questions*, *Mental Effort*, *Self-Explanations*, and *Time*, the two conditions were compared with *t*-tests. Regression analyses were done to establish the indirect effect of *Guidance* via *Mental Effort* and *Self-Explanations* on both *Achievement Questions* and *Transfer Questions* separately. In both regression analyses *Time* was included in the model as a covariate together with the interaction term *Mental Effort* x *Self-Explanations*.

The second hypothesis was that guidance would raise the efficiency of student learning. To confirm this hypothesis a *t*-test was done to compare the *Efficiency* of the two conditions.

The third hypothesis was that guidance via performance would lead to a more positive attitude. To confirm this hypothesis again a *t*-test and a supplementary regression analysis were done. To test for the direct effect of *Guidance* on *Attitude* a *t*-test was done. To establish the effect of *Guidance* via *Performance* on *Attitude* a regression analysis was done.

To determine the construct validity of the attitude questionnaire the correlations were calculated among the four separate attitudinal aspects (*Value, Difficulty, Affect, Cognitive Competence*) and *Performance*. Cronbach's *α* was computed for the *Achievement Questions*, *Mental Effort*, and the four constructs of the attitude questionnaire. Interrater reliability with two raters was measured of the *Achievement Questions* and the *Transfer Questions*. Mean scores and standard deviations of the raters and a correlation between the scores were computed. As a randomisation check *Grades* of the students in both conditions were compared with a *t*-test.

### 6.13 Results

With regard to the first hypothesis it was expected that students in the intervention condition would perform better, perceive a lower cognitive load and report more correct self-explanations. The *t*-tests showed significant differences on *Achievement Questions*, *Transfer Questions*, and *Self-Explanations*, but no significant difference on *Mental Effort*, and *Time*.

Thus, students in the intervention condition performed better and reported more correct self-explanations, but perceived the same cognitive load as the control group. The results of the *t*-tests are presented in Table 6.1.

Table 6.1. Results of the *t*-tests comparing the intervention and control condition

| | $M_{control}$ | $M_{intervention}$ | *S.E.* | *t* | *p*-value |
|---|---|---|---|---|---|
| *Achievement Questions* | 5.40 | 9.71 | 1.16 | 3.72 | .002 |
| *Transfer Questions* | 3.96 | 4.75 | .52 | 3.42 | .002 |
| *Self-explanations* | 5.76 | 19.21 | 1.66 | 8.19 | .000 |
| *Mental Effort* | 64.88 | 66.88 | 2.56 | .78 | .876 |
| *Efficiency* | -.38 | .49 | .29 | 3.04 | .004 |
| *Attitude* | 80.48 | 86.46 | 4.83 | 1.24 | .222 |
| *Time* | 64.23 | 67.42 | 2.55 | 1.25 | .567 |
| *Grades* | 7.62 | 7.75 | .47 | .27 | .792 |

Note. *p*-values Bonferroni corrected.

The regression analysis of *Achievement Questions* on *Guidance, Mental Effort*, and *Self-Explanations* showed a significant effect of *Mental Effort* and *Self-Explanations,* but no significant effect of *Guidance*. The covariates *Time* and the interaction term *Mental Effort* x *Self-Explanations* had no significant effect and were excluded from the model in a backward procedure. The results of the final regression model are presented in Table 6.2.

Table 6.2. Regression of *Achievement Questions* on *Guidance*, *Mental Effort* and *Self-Explanations*

| *Model* | *B* | *S.E* | *t* | *p*-value |
|---|---|---|---|---|
| *Constant* | 10.00 | | | |
| *Guidance* | -1.29 | 1.34 | - 0.97 | .337 |
| *Mental Effort* | -0.11 | 0.05 | - 2.24 | .030 |
| *Self-Explanations* | 0.43 | 0.08 | 5.68 | .000 |

Note: The non-significant covariates *Time* and *Mental Effort* x *Self-Explanations* were excluded from the model in a backward stepwise model-selection procedure.

The regression analysis of *Transfer Questions* on *Guidance, Mental Effort*, and *Self-Explanations* analogously showed a significant effect of *Mental Effort* and *Self-Explanations,* but no significant effect of *Guidance*, *Time* and the interaction term *Mental Effort* x *Self-Explanations*. The results of this analysis are presented in Table 6.3.

These results support the first hypothesis that the effect of *Guidance* is manifested via *Mental Effort* and *Self-Explanations*.

Table 6.3. Regression of *Transfer Questions* on *Guidance*, *Mental Effort* and *Self-Explanations*

| *Model* | *B* | *S.E* | *t* | *p*-value |
|---|---|---|---|---|
| *Constant* | 9.47 | | | |
| *Guidance* | 0.37 | 0.58 | 0.63 | .533 |
| *Mental Effort* | -0.11 | 0.02 | -5.22 | .000 |
| *Self-Explanations* | 0.12 | 0.03 | 3.70 | .000 |

Note: The non-significant covariates *Time* and *Mental Effort* x *Self-Explanations* were excluded from the model in a backward stepwise model-selection procedure.

The second hypothesis with respect to the effect of *Guidance* on *Efficiency* was confirmed. *Efficiency* was significantly higher for students in the intervention condition (see Table 6.1).

The third hypothesis referred to the attitude toward statistics. It was hypothesised that the students who perform better due to guidance, would display a more positive attitude toward statistics after the experiment.

This hypothesis was not confirmed. The *t*-test comparing the A*ttitude* of the two conditions showed no significant difference (see Table 6.1). The regression of *Attitude* on *Performance* and *Guidance* showed only a significant effect of *Performance*. The results of this analysis are presented in Table 6.4.

Table 6.4. Regression of *Attitude* on *Guidance* and *Performance*

| *Model* | *B* | *S.E* | *t* | *p*-value |
|---|---|---|---|---|
| *Constant* | 63.63 | | | |
| *Guidance* | -6.32 | 4.35 | -1.46 | .154 |
| *Performance* | 2.02 | 0.36 | 5.61 | .000 |

The correlations among the subscales of the attitude questionnaire and *Performance* are presented in Table 6.5. The pattern of correlations is in line with the results reported by Gal, et al. (1997) and Budé, et al (2006). This seems to indicate that our modification of the SATS has proper construct validity.

Reliability analyses showed that *Achievement Questions* ($\alpha$ = .72), *Transfer Questions* ($\alpha$ = .68), and *Mental Effort* ($\alpha$ = .73), as well as the four subscales of the attitude questionnaire: *Value* ($\alpha$ = .74), *Difficulty* ($\alpha$ = .76), *Affect* ($\alpha$ = .73), and *Cognitive Competence* ($\alpha$ = .74), were quite reliable.

Table 6.5. Correlations among *Value*, *Difficulty*, *Affect*, *Cognitive Competence* and *Performance*

| | *Value* | *Difficulty* | *Affect* | *Cognitive Competence* |
|---|---|---|---|---|
| *Value* | 1 | .362 | .691* | .484* |
| *Difficulty* | | 1 | .646* | .611* |
| *Affect* | | | 1 | .701* |
| *Cognitive Competence* | | | | 1 |
| *Performance* | .455* | .450* | .650* | .514* |

Notes. * $p < .05$; $p$-values Bonferroni corrected. $N = 49$.

Interrater reliability of *Achievement Questions* and *Transfer Questions* was high. The correlation between the question scores was high ($N = 10$; $r = .955$), the mean ratings did not significantly differ ($M_{\text{rater 1}} = 14.7$; $M_{\text{rater 2}} = 13.8$; $t = 1.13$, $p = .287$), and the standard deviations were equal ($SD_{\text{rater 1}} = 7.9$; $SD_{\text{rater 2}} = 6.7$).

The randomisation check showed no difference between the intervention and the control condition. *Grades* of students in both conditions did not differ significantly (see Table 6.1).

## 6.13 Discussion

In this study the positive aspects of human tutoring and learning with worked examples were combined in guidance by directive questions. It was expected that this form of guidance would stimulate self-explanations, decrease cognitive load, and enhance performance of statistics. The results of this study show that the directive questions stimulated self-explanations and enhanced performance on the achievement questions, without raising the cognitive load. Consequently the efficiency of the students in the intervention condition was significantly higher than in the control condition.

Moreover, the efficiency in the control condition had a negative sign. This means that the standardised scores for cognitive load were higher than those for performance. In contrast, in the intervention condition the standardised performance was higher than the standardised mental effort. This seems to indicate that providing directive questions is an efficient way to successively guide the students' line of reasoning, which in turn enhances the performance without raising cognitive load.

The direct effect of guidance, as shown by the significant results of the *t*-tests on the achievement and transfer questions, disappeared in the regression analyses when self-explanations and cognitive load were included as mediating variables in the model. This indicates that the effect of guidance is mediated by these two factors (Baron, & Kenny, 1986). In other words, the effect of guidance is manifested through stimulated self-explanations and reduced cognitive load as we hypothesised. Time-on-task had no significant effect, neither when it was examined in the *t*-test, nor in the regression analyses. This means that the time spent on answering the questions had no confounding effect in this study.

The achievement questions were designed to elicit from the students the explanation of the relations between several statistical concepts and the applicability of these concepts. Answering such questions correctly is only possible when a higher level of understanding is attained (Dochy, 2001; Feltovich, et al., 1993; Gijbels, et al., 2005; Jonassen, et al., 1993). Students in the intervention condition showed a better performance on the achievement questions. Moreover, they also answered the transfer questions significantly better than the students in the control condition. As transfer questions are regarded as the ultimate measure of higher levels of understanding (Barnett & Ceci, 2002; Catrambone, 1998; Mayer, 1989; Olson & Biolsi, 1991), the enhanced performance on both the achievement and the transfer questions can be interpreted as a higher level of understanding of the students in the intervention condition.

The attitude toward statistics was hypothesised to be more positive for students in the intervention condition. The attitude questionnaire consisted of statements on studying statistics in general as well as on the participation in the experiment. Although there was a positive trend in the attitude, this effect did not reach significance. We did find, however, a significant effect of performance on attitude. This means that, aggregated over the two conditions, students who performed better showed a more positive attitude. This is in line with our hypothesis, but the guidance effect of the intervention was probably too small to reach significance.

An additional problem is that it cannot be ruled out that the students who volunteered to participate in the study had a more positive attitude than the overall population of students. In this light it is worthwhile to consider the mean grades of the preceding statistics course. The mean grades are rather high for the students in both conditions. Possibly, relatively good students with already a positive attitude prior to the study volunteered to participate. So, it can only be tentatively concluded that performance is positively correlated with attitude. This is in line with theories regarding attitudes (Graham & Weiner, 1987; Peterson, et al., 1993; Pintrich & Schunk, 1996; Weiner, 1986, 1992). Of the four attitude scales, affect was most strongly correlated with performance. This corroborates previous research on attitudes and performance (Budé, et al., 2006). Stimulating a positive affect of students toward statistics, may improve performance in statistics education.

The results also showed that the reliability of all the measurements was relatively high. Moreover, the reliability of the subscales of the attitude questionnaire as well as the pattern of correlations among them is in line with previous research. We found no significant correlation between value and difficulty, the strongest correlation between affect and cognitive competence, and moderate correlations for the other combinations. This resembles the pattern of the SATS that is reported by Gal, et al. (1997). The similarities of these patterns can be interpreted as a validation of the modification of the SATS that was used in this study.

A limitation of this study is that possibly more proficient students volunteered to participate. The randomisation check showed no significant

difference between the two conditions regarding the mean grades on the exam of the preceding statistics course. As mentioned above, however, the mean grades of students in both conditions were higher than the mean grade of the rest of the cohort of bachelor students. This might indicate that the students in the study had more prior knowledge of the subject matter than the overall population of students. It can not be ruled out that directing reasoning processes by asking guiding questions, requires a certain level of prior knowledge of the students. In future research it could be investigated whether the guiding questions have the same positive effect on students' understanding with different levels of prior knowledge.

A second limitation of the study is that although students in the control condition were stimulated to reflect on their responses and asked to write down their reasoning steps, this treatment may not have activated them as much as the students in the intervention condition. Future research could be directed at a comparison between guiding questions and questions that are less directive for reasoning activities. In this way mere activation of reflection can be contrasted with directing the reasoning processes as we did in our study.

In the present study guidance was supposed to have an opposite effect on cognitive load and self explanations. Cognitive load was supposed to be lowered and self-explanations were expected to be enhanced. Due to the recursive relation between cognitive load and self-explanations it could not be predicted what the resulting level of either of them separately would be. We included the interaction of cognitive load and self-explanations in the regression analyses. In neither of the analyses this interaction was significant. We acknowledge that an interaction term is not the most suited way to model a recursive relation, but the design of the present study did not enable a further investigation of the recursiveness of this relation.

In conclusion, without giving the students specific information about the correct answers on the achievement questions, i.e. by providing directive questions, students in the intervention condition were stimulated to more actively apply and expand their knowledge, and this resulted in a higher level of understanding. It can be inferred that this higher level of understanding was caused by the guiding questions in this randomised experiment, because it can be presumed that before the start of the study, students in both conditions had an equal understanding of the topics that were questioned.

The conclusion that directive guidance leads to better understanding may have wide implications for educational practice. Directive questions that step by step guide the correct activation of students' prior knowledge could for example be used in problem based learning, in textbooks, in formative testing, or in electronic learning environments.

## 6.15 Appendix A:

*A hypothetical study and examples of the achievement questions.*
A health scientist studied the effect of a drug in two randomised groups of patients with hypertension. One group received a placebo (condition = 0); the other group received the drug for one month (condition = 1). After this month the researcher measured the diastolic blood pressure of all subjects and compared the two groups. The research hypothesis was: the mean diastolic blood pressure of the group patients who received the drug will be lower than the mean diastolic blood pressure of the group patients who received the placebo. The data were analysed with a *t*-test, linear regression analysis, and analysis of variance. The results of these analyses are presented below.

## T-Test

**Group Statistics**

| | CONDITION | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| BLOOD PRESSURE | 0 | 30 | 98.367 | 9.6888 | 1.7689 |
| | 1 | 30 | 93.043 | 8.5685 | 1.5644 |

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| | | | | | | | | | Lower | Upper |
| Blood pressure | Equal variances assumed | .111 | .740 | 2.255 | 58 | .028 | 5.324 | 2.3615 | .5972 | 10.0511 |
| | Equal variances not assumed | | | 2.255 | 57.146 | .028 | 5.324 | 2.3615 | .5957 | 10.0526 |

## Regression

**Variables Entered/Removed(b)**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | CONDITION(a) | . | Enter |

a All requested variables entered.
b Dependent Variable: BLOOD PRESSURE

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .284(a) | .081 | .065 | 9.1459 |

a Predictors: (Constant), CONDITION

**ANOVA(b)**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 425.198 | 1 | 425.198 | 5.083 | .028(a) |
| | Residual | 4851.512 | 58 | 83.647 | | |
| | Total | 5276.709 | 59 | | | |

a Predictors: (Constant), CONDITION
b Dependent Variable: BLOOD PRESSURE

**Coefficients(a)**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | | 95% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. | Lower Bound | Upper Bound |
| 1 | (Constant) | 98.367 | 1.670 | | 58.909 | .000 | 95.024 | 101.709 |
| | CONDITION | -5.324 | 2.361 | -.284 | -2.255 | .028 | -10.051 | -.597 |

a Dependent Variable: BLOOD PRESSURE

## One-way ANOVA

**ANOVA**

BLOOD PRESSURE

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between groups | 425.198 | 1 | 425.198 | 5.083 | .028 |
| Within groups | 4851.512 | 58 | 83.647 | | |
| Total | 5276.709 | 59 | | | |

### Examples of achievement questions

**Question 1.** How would you interpret the result of the *t*-test? *Explain* your conclusion with regard to the research hypothesis.
**Question 2.** What is the value of the regression coefficient for the independent variable? *Explain* what effect this coefficient represents.
**Question 3.** In the regression analysis table a *t*-test is given for condition. Describe and *explain* what is tested with this test and why.
**Question 4.** Write down the regression equation and calculate the mean diastolic blood pressure of both groups with this equation.
**Question 5.** In the ANOVA-table an *F*-value is given. Describe how it is calculated and *explain* why this is done in this way (*explain* the rationale for this procedure).

### Examples of guiding questions (pertaining to the achievement questions above).

**Guiding questions for achievement question 1:**
How is a *t*-value calculated, what formula is used?
What is the *t*-value here, what is its p-value?

When is a null hypothesis rejected?
What does the *t*-test say about the two groups in the presented experiment?

**Guiding questions for achievement question 2:**

Plot the regression line in a x- and y-axis diagram
What is the independent variable, what is the dependent variable?
Through which points does the regression line go?
What do regression coefficients stand for?
What does the coefficient for the independent variable in this analysis stand for?

**Guiding questions for achievement question 3:**

If there is no difference between the two groups, what would the slope of the regression line be?
What value would the regression coefficient of the independent variable have?
What test can be done to establish whether there is a difference between the groups?

**Guiding questions for achievement question 4:**

Consider again the regression line that you drew. What/where is the intercept?
What does the intercept stand for in this case?
Where did you pinpoint the two conditions?

**Guiding questions for achievement question 5:**

What information is used in the analysis of variance?
What sources can you distinguish and what do they mean?
How is the *F*-ratio constructed?
What is in the numerator, what is in the denominator of the *F*-ratio?

## 6.16 Appendix B

*An example of a transfer question and the alternative hypothetical study.*
A dermatologist studied the effect of an ointment in two randomised groups of patients with a skin disease. One group received the ointment for one month; the other group received a placebo. After this month the researcher measured the number of complaints of all subjects and compared the two groups. The research hypothesis was: the number of complaints of patients who received the ointment will be lower than of patients who received the placebo. The data were analysed with linear regression.
Interpret the results, in terms of means, $R^2$-value, *p*-value, research hypothesis, etc, of these analyses below as completely as possible. *Explain* your conclusions.

**Variables Entered/Removed(b)**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | CONDITIE(a) | . | Enter |

a All requested variables entered.
b Dependent Variable: ECZEEM

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .131(a) | .017 | .000 | 17.5263 |

a Predictors: (Constant), CONDITIE

**ANOVA(b)**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 312.209 | 1 | 312.209 | 1.016 | .318(a) |
| | Residual | 17815.974 | 58 | 307.172 | | |
| | Total | 18128.184 | 59 | | | |

a Predictors: (Constant), CONDITIE
b Dependent Variable: ECZEEM

**Coefficients(a)**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | 23.977 | 3.200 | | 7.493 | .000 | 17.572 | 30.383 |
| | CONDITIE | -4.562 | 4.525 | -.131 | -1.008 | .318 | -13.621 | 4.496 |

a Dependent Variable: ECZEEM

## 6.17 Appendix C

*The modified SATS, per subscale. The modified or added statements are marked with an asterisk. The original statements were translated into Dutch. The statements were presented in a random order. Each of the items had to be rated on a 7-point Likert-scale.*

**Value**
Statistics is worthless.
Participating in this study was useful*.
I consider other courses to be more interesting.
Statistical thinking is not applicable in my everyday life.
Statistics is irrelevant in my life.
The questions in this study were interesting*.
The questions in this study were relevant for upcoming statistics courses*.

**Difficulty**
Statistics is easy to understand.
Statistics is a complicated subject.
It was easy to answer the questions of this study*.
Statistics is a subject quickly learned by most people.

**Affect**
I like statistics
I enjoy taking statistics courses.
I did not like to participate in this study*.
I feel insecure when I have to do statistics problems.
I was under stress during this study*.
I am scared by statistics.

**Cognitive competence**
I have trouble understanding statistics because of how I think.
I had no idea what the questions in this study were about*.
I feel confident that I will pass all my statistics exams*.
I can learn statistics.

# Chapter 7

## Summary and general conclusions

## 7.1 Goal

The main purpose of this dissertation was to investigate interventions in statistics education of health sciences' students. Students' conceptual understanding of statistics in the social and health sciences is usually quit low. The aim of the interventions was to improve conceptual understanding of these students. Students' understanding of the subject matter is to a large extent determined by their learning. How and what students learn is partially determined by their motivation and by education. Improved education should, therefore, lead to better conceptual understanding and hence to better performance. In the figure that is presented in chapter 1 the relations between education, motivation, learning, conceptual understanding, and performance are depicted.

These constructs and their relations were investigated in a series of studies. In chapter 2 a method for the measurement of conceptual understanding was studied. In chapter 3 a motivational model was tested in the domain of statistics education. In chapters 4 and 5 the effects of interventions in statistics education on three different outcome variables were studied. In chapter 4 the effects of more directive tutor guidance on students' performance and course evaluation were studied. In chapter 5 the effect of the distribution of practice and directive tutor guidance on student's conceptual understanding was investigated. In chapter 6 the effects of written directive questions 1) on two learning processes: cognitive load and reasoning steps, 2) on conceptual understanding, and 3) on motivation, were studied in an experiment.

## 7.2 The studies

### *7.2.1 Study 1*

In **Chapter 2** a method for the assessment of conceptual understanding was studied. We defined conceptual understanding as knowing all the relevant concepts as well as the relations between these concepts. From this definition it follows that conceptual understanding is related to the knowledge structures of an individual. Based on this relation three criteria for the assessment of conceptual understanding were formulated. These criteria pertain to properties of the knowledge structures.

The three criteria are: 1) the richness of these knowledge structures, 2) the integration of new knowledge into existing knowledge structures, and 3) the presence of not explicitly presented knowledge. These criteria were used to develop a method to measure conceptual understanding. This assessment method was used in a study of conceptual understanding of a science text. First, the method consisted of two kinds of open ended questions: questions that directly tapped on presented knowledge and questions that asked for transfer of knowledge. All questions asked explicitly from the participants to explain in their own words the relations between the science concepts. The use of all the open ended questions thus revealed the *richness* and the *integration* of the knowledge. Transfer questions in particular asked for knowledge that was *not explicitly presented* during the learning phase.

Second, the assessment method consisted of four scoring techniques for the answers to all the questions. These four scoring techniques were:

- A classical scoring method, which focussed on a global intuitive assessment of the answers by two science teachers.
- A scoring method with an elaborate answer key. This method focussed solely on the number of mentioned science concepts.
- A scoring method in which the written answers were first transformed into semantic networks, which were subsequently assessed. This method focussed on the number and content of the mentioned concepts as well as on the relations between them.
- A scoring method that assessed pathfinder networks, which were constructed on the basis of the written answers. The focus of this scoring method was on the relatedness of the concepts.

Together the four complementary scoring techniques enabled the assessment of the *richness* of the knowledge structures, the level of *integration* of newly acquired knowledge into prior knowledge, and the existence of *not explicitly presented* knowledge.

Results show that this assessment approach seems to be a useful method. The method seems to be a practical, reliable, and valid way to measure what learners understand from subject matter. It is concluded that using an answer key might be the most practical method for scoring the written answers, because it is the least laborious method. In the studies of chapter 5 and 6 this method was used for the measurement of conceptual understanding of statistics.

*7.2.2 Study 2*

In **Chapter 3** the relation between a number of motivational constructs and performance was studied. The motivational constructs that we chose to study were:

- Causal explanations of statistics related events, for example, perceived causes for failing the statistics exams.
- Perceived outcome expectancy of students' activities within a statistics course, for example, the profit the student expects from attending a lecture.
- Affect, for example, interest and enjoyment.
- Study behaviour toward statistics, for example, effort and persistence.

The mutual relations of these constructs, as well as their effect on the performance on the course exam, were studied in the domain of statistics education. Students, participating in an introductory statistics course, filled in a questionnaire that measures these motivational constructs. A confirmative factor analysis and a path analysis were done with Lisrel.

The first finding concerns the stability of and control over causes of statistics related events. It was shown that students who think that they lack *control* over, for example, causes for failing the statistics exams, did not expect to profit from studying statistics. Students who think that there are *stable* causes for failing, but in spite of that do invest time, started to dislike statistics. A second finding concerns the importance of students' affect. Students who appreciate the value and relevance of statistics, who think it is interesting, challenging, and who like statistics, appear to study statistics more and qualitatively better, and perform better on the exams. The third main finding in this study concerns study

behaviour. Two aspects of study behaviour were measured in this study, effort and persistence. Effort is operationalised as the endeavour of the students, i.e. the amount of time they spent on studying statistics, whether they attended the lectures, and whether they prepared themselves for the tutorial meetings. Persistence is operationalised as the tenacity of the students, i.e. whether students turned to their lecture notes or their books, or consulted a teacher when they were not able to solve a statistical problem. It was shown that, more than just mere effort, persistence in studying statistics contributes significantly to a positive result on the exam. This result suggests that persisting in trying to find the solution to a difficult topic, instead of switching over to another topic, is the best way to study statistics.

*7.2.3 Study 3*

In **Chapter 4** the effects of directive tutor guidance in problem based learning (PBL) of statistics on students' performance and on the course evaluation were studied. Two conditions were compared. In both conditions statistics was taught in a PBL educational setting, with tutorial meetings. In the first condition the tutors had a more traditional role during the discussions in the meetings, i.e. they tried to activate students, stimulated group processes, tried to create an atmosphere in which students could optimally participate in the discussions, helped students to monitor their own learning, and stimulated self-study. In the intervention condition, the tutors guided the discussions in a more directive way. This directive tutor guidance consisted of questions the tutors used to structure the discussions and to indicate the way of reasoning during the tutorial meetings. The effects of this directive tutor guidance on students' performance and on a number of affective aspects toward the statistics course are studied.

The results showed that performance improved marginally significant and that the course was valued significantly more positive. Directive tutor guidance, therefore, seems to be an effective improvement in problem based learning (PBL) of statistics.

*7.2.4 Study 4, 5, and 6*

In **Chapter 5** the effect of the distribution of practice and directive tutor guidance on student's conceptual understanding was investigated in three consecutive studies. Conceptual understanding was measured in these three studies with the same instrument. The instrument that was designed for these studies consisted of open ended questions. In the first study a six months' statistics course was compared to a more condensed course of eight weeks. The two courses were comparable with regard to the content and were both PBL courses. However, in the condensed course the students were forced to more massed practice, i.e. studying subject matter uninterruptedly or with only short breaks, in a brief time interval.

The results show that massed practice negatively affected students' conceptual understanding. In the second study the effect of directive tutor guidance on conceptual understanding was investigated. In the guidance

condition the tutors, as in the study of chapter 4, asked directive questions to structure the discussions. We found that directive tutor guidance leads to improved conceptual understanding. This supported the findings from the study in chapter 4, where an improved performance on the final course exams was found. In the third study long term retention was studied. We found that long term retention six months after the course improves by directive tutor guidance and by distributed practice. Yet, students' conceptual understanding after six months did not show much improvement compared to the level of conceptual understanding at the entry test.

*7.2.5 Study 7*

In **Chapter 6** the effect of guidance by directive questions on students' conceptual understanding and their attitude was investigated in an experiment. In the field studies from chapters 4 and 5 tutors asked questions during the discussions in the tutorial meetings and it was shown that this led to improved performance. However, due to the differences between the tutors and due to differences in the discussions, the actual questions the tutors asked could not be controlled in those studies. In this experimental study standardised written questions were used. As a result the intervention was equal for all students, i.e. possible dissimilarities in the intervention of the field studies were cancelled out. The goal of this study was, first, to confirm the results of the field studies in an experiment. A second goal was to study two to student learning related mechanisms, that is, self-explanations and cognitive load, as possible mediators. The third goal was to study the effect on students' attitude.

The results of this experimental study show that the standardised directive questions led to better conceptual understanding as directive tutor guidance did. The results also show that this positive effect of the directive questions was mediated by self-explanations and cognitive load. Providing directive questions stimulated self-explanations without raising the overall cognitive load, that is, the sum of extraneous and germane cognitive load. This means that the reasoning processes of the students were more efficient, which led to improved conceptual understanding. So, the results not only corroborate the findings from the studies in chapters 4 and 5, they also reveal a possible mechanism that explains the positive effect of asking questions on students' conceptual understanding.

An effect of guidance by directive questions on students' attitude was not found. We did find that students who performed better showed a more positive attitude toward statistics. Based on this result, it could only tentatively be concluded that performance is positively correlated with attitude.

## 7.3 General conclusions

Several conclusions can be drawn from the studies in this dissertation. First, guidance in statistics education through directive questions enhances students' conceptual understanding. In both studies in which the tutors asked questions to direct the discussions in the tutorial meetings, the students in the intervention condition did better than in the control condition. In the study from chapter 4 they performed better on the exam. In the study from chapter 5 they

showed improved conceptual understanding as measured with the open ended questions. Moreover, the experiment in chapter 6, in which written directive questions were used, shows analogously a positive effect on students' conceptual understanding. Guiding students through the subject matter and directing their reasoning processes supported students' learning in two ways. First, through guidance their attention was focussed on the important aspects of the subject matter. In other words, the directive questions indicated the students *how* to deal with the subject matter. This made their reasoning more accurate, and more focussed on relevant subjects. Second, the directive questions also forced the students to elaborate on difficult topics. Hence, the questions also activated students' reasoning processes. The study in chapter 6 shows that although students in both conditions were encouraged to reflect on their reasoning, students in the intervention condition reported significantly more reasoning steps than students in the control condition. Apparently it seems not enough to create a facilitating context; students also have to be activated. It is well accepted in educational psychology that active learning improves conceptual understanding of the subject matter. So, there are two ways in which directive questions improved conceptual understanding. First, the directive questions indicated to the students how to deal with the subject matter. Second, they activated the students. When these two factors are combined, it can be concluded that directive questions made students' reasoning more effective. In terms of cognitive load theory, extraneous load was reduced and germane load was increased.

A conclusion that can be drawn from the study in chapter 2 is that conceptual understanding reliably and validly can be measured with the open ended questions that were used in the study. Asking for the relations between concepts reveals the richness of an individual's knowledge structures. Formulating the answers to the questions in ones own words and giving self contrived examples reveals the level of integration of new acquired knowledge. Transfer questions ask for knowledge that was not presented during the learning phase. The questions thus reveal those properties of an individual's knowledge structures that we have defined as being characteristic for conceptual understanding. The study also shows that in practice it suffices to score the answers with an answer key.

A third conclusion that can be drawn from the studies in this dissertation is that students' affect toward statistics is important. Affect not only seems to be directly related to performance, but also indirectly via persistence in study behaviour. In the study in chapter 3 it has been shown that persistence leads to improved performance. Although affect seems to be a relevant point of interest, innovations in statistics education are rarely specifically directed at making the courses more attractive, interesting, or pleasant. Stimulating persistence in studying statistics per se, is a theme that needs more attention as well.

A fourth conclusion is that distributed practice enhances conceptual understanding. In the Netherlands there is a trend to organise education in courses with a relatively short time span. This trend can be seen at universities, high schools, and even in secondary schools throughout the Netherlands. It

should be noted that in educational systems with such brief courses, the possibility for distributed practice is limited. The result of the study in chapter 5 indicates that the restriction of distributed practice may have negative consequences for the level of conceptual understanding that students can achieve.

Finally, it can be concluded that students' conceptual understanding and long term retention of statistics seems to be rather low in the social and health sciences. In the studies from chapters 5 and 6 open ended questions were used to measure conceptual understanding. The results show that the answers to the open ended questions were rather inaccurate and the level of long term conceptual understanding was not much improved compared to the entry test. This fact also calls for attention in the development and improvement of statistics education.

### 7.4 Remarks

In this dissertation five field studies are described. In these studies, although groups were randomised, we could not control all factors, which may have reduced the internal validity of the studies to some extent. On the other hand, because the studied conditions were embedded in practical field situations, we think that the external validity of these studies may be rather high. The study in chapter 6 was an experimental study. Because the results from this study concur with the results from the field studies, it may be inferred that our conclusions regarding directive guidance are both externally as internally valid. However, further research is needed to investigate how the directive questions of the tutors affect the discussions in the meetings, whether asking directive questions also has a positive effect on more proficient students, and how long term retention can be improved. Future research could also be directed at the recursive relation between cognitive load and self-explanations.

### 7.5 Recommendations

Instruction should be rather directive, students should not get lost in subject matter that is new to them. Asking directive questions seems to be a very effective instruction tool in the teaching of statistics. Directive questions guide the students through the topics and point them into the correct direction, instead of giving the solution to a problem promptly. This might stimulate both students' activity and their persistence. Moreover, when students are able to successfully answer questions, this might lead to a more positive affect toward statistics.

Students' affect toward statistics should be a point of special interest in statistics education. It should be tried to make the courses more attractive, interesting, and enjoyable for the students.

Statistics courses should also encourage students' active participation and stimulate their motivation by giving the students the opportunity to experience success. If students experience success they will expect more profit from studying statistics. This may lead to more active participation in the courses. Experiencing success might also lead to a more positive affect. In practice this means that in constructing a learning environment, there should be tasks built in

that students can accomplish. More difficult tasks should only be administered after students have attained a more positive motivation toward statistics.

To increase the possibility of distributed practice courses should be spread out more in time, or subject matter should recur after the course is terminated. The integration of statistics in other content oriented courses, could bring about such a recurrence. Without distributed practice or a periodic recurrence of the subject matter, the relations between concepts will fade and conceptual understanding will decline. Improving statistics education should not only be directed at improving short-term conceptual understanding, but also at the maintenance of it.

## References


Albanese, M.A., & Mitchell, S. (1993). Problem-based learning: A review of literature on its outcomes and implementation issues. *Academic Medicine, 68,* 52-81.

Ames, C. (1992). Classrooms: goals, structures, and student motivation. *Journal of Educational Psychology*, *84*, 261-271.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge: Harvard University Press.

Anderson, J. R., & Schunn, C. D. (2000). Implications of the ACT-R learning theory: No magic bullets. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science*. Mahwah: Lawrence Erlbaum Associates, Publishers.

Baddeley, A. (1992). Working memory, *Science, 255*, 556-559.

Baddeley, A.D., (2000). Short term and working memory. In E. Tulving and F.I.M. Craik (Eds.), *The Oxford handbook of memory*, New York: Oxford University Press.

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52*, 566-577.

Bahrick, H. P., & Phelps, E. (1987). Retention of Spanish vocabulary over eight years. *Journal of Experimental Psychology, 13*, 344-349.

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the learning Sciences, 13*, 1-14.

Barnett, S.M., & Ceci, S.J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*, 612-637.

Baron, R. M., Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Barrows, H. S. (1986). A taxonomy of problem-based learning methods. *Medical Education, 20*, 481-486.

Barrows, H. S. (1988). *The Tutorial Process*. Springfield: Southern Illinois University School of Medicine.

Bassok, M. (1997). Object based reasoning. In D. L. Medin (Ed.), T*he Psychology of learning and motivation*. San Diego, Academic Press.

Bassok, M. (2003). Analogical transfer in problem solving. In J. E. Davidson, & R. J. Sternberg (Eds.), *The psychology of problem solving*. Cambridge: Cambridge University Press.

Benjamin, A. S., & Bird, R.D. (2006). Metacognitive control of the spacing of study repetitions. *Journal of Memory and Language, 55*, 126-137.

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

Biggs, J.B. (1993). What do inventories of students' learning processes really measure? A theoretical review and clarification. *British Journal of Educational Psychology, 63*, 3-19.

Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction, 7*, 161-186.

Broers, N. J. (2006, July). *Learning goals: The primacy of statistical knowledge.* Paper presented at the International Conference on Teaching of Statistics, Salvador, Bahia.

Bruning, R.H., Schraw, G.J., Ronning, R.R. (1999). *Cognitive psychology and instruction.* Upper Saddle River, NY: Merrill, Prentice Hall.

Budé, L. (2006, July). *Assessing students' understanding of statistics*. Paper presented at the International Conference on Teaching of Statistics, Salvador, Bahia.

Budé, L., Imbos, Tj., van de Wiel, M., M.J.J.M., Broers, N.J., & Berger, M.P.F. (2005, August). *Tutor guidance in problem-based learning*. Paper presented at biannual conference of the European Association for Research on Learning and Instruction, Nicosia, Cyprus.

Budé, L., van de Wiel, M., Imbos, Tj., Candel, M.J.J.M., Broers, N.J., & Berger, M.P.F. (2006). Students' achievements in a statistics course in relation to motivational aspects and study behaviour. Manuscript submitted for publication.

Budé, L. M., van de Wiel, W. J., Imbos, Tj., Schmidt, H. G., & Berger, M. P. F. (2006). *A procedure for measuring understanding*. Manuscript submitted for publication.

Campione J. C. & Brown, A. L. (1990). Guided learning and transfer: Implications for approaches to assessment. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition*. Hillsdale: Lawrence Erlbaum Associates, Publishers.

Carroll, W.M. (1994). Using worked examples as an instructional support in the algebra classroom. *Journal of Educational Psychology,86,* 360-367.

Catrambone, R. (1998). The sub goal learning model: Creating better examples so that students can solve novel problems. *Journal of experimental psychology, 127*, 355-376.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, 10(3).

Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology, 10*, S33-S49.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reiman, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*, 145- 182.

Chi, M.T.H., de Leeuw, N., Chiu, M-H., & LaVanger, C. (1994). Eliciting self-explanations improves learning. *Cognitive science, 18*, 439-477.

Chi, M. T., Feltovitch, P. J. & Glaser, R. (1981). Categorisation and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Chi, M. T., Siler, S.A., Jeong, H., Yamauchi, T., & Hausman, R.G. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471-533.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82,* 407-428.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8,* 240-248.

Conway, M. A., Cohen, G., & Stanhope, N. (1991). On the very long-term retention of knowledge acquired through formal education: Twelve years of cognitive psychology. *Journal of Experimental Psychology, 120*, 395-409.

Conway, M. A., Cohen, G., & Stanhope, N. (1992). Very long-term retention of knowledge acquired at school and university. *Applied Cognitive Psychology, 6*, 467-482.

Cooke, N. J. (1992). Predicting judgment time from measures of psychological proximity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*, 640-653.

Cooke, N. J., Durso, F. T. & Schvaneveldt, R. W. (1986). Recall and measures of memory organisation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *12,* 538-549.

Dearholt, D. W., & Schvaneveldt, R., W. (1990). Properties of Pathfinder Networks. In R., W. Schvaneveldt (Ed), *Pathfinder associative networks. Studies in knowledge organization* (p. 1-30). Norwood, Ablex Publishing Corporation.

Debowski, S, Wood, R.E., & Bandura, A. (2001). Impact of guided exploration and enactive exploration on Self-regulatory mechanisms and information acquisition through electronic search. *Journal of Applied Psychology, 86,* 1129-1141.

Deci, E.L., & Ryan, R.M. (1985). *Intrinsic motivation and self-determination in human behavior.* New York: Plenum Press.

De Grave, W. S., Dolmans, D. H. J. M., & Van Der Vleuten, C. P. M. (1999). Profiles of effective tutors in problem-based learning: scaffolding student learning. *Medical Education, 33*, 901-906.

De Jong, T., & Ferguson-Hessler, M.G.M (1996). Types and qualities of knowledge. *Educational Psychologist, 31*, 105-113.

delMas, R. C. (2002). Statistical literacy, reasoning and learning: a commentary. *Journal of Statistics Education,* 10(3).

delMas, R. C., Garfield, J., & Chance, B. L. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education, 7*.

Dochy, F. (2001). A new assessment era: different needs, new challenges. *Research Dialogue in Learning and Instruction, 2*, 11-20.

Dolmans, D. H. J. M., Gijselaers, W. H., Moust, J. H. C., De Grave, W. S., Wolfhagen, I. H. A. P., & Van Der Vleuten, C. P. M. (2002). Trends in research on the tutor in problem-based learning: conclusions and implications for educational practice and research. *Medical Teacher, 24*, 173-180.

Dolmans, D. H. J. M. & Wolfhagen, I. H. A. P. (2005). Complex interactions between tutor performance, tutorial group productivity, and the

effectiveness of PBL Units as perceived by students. *Advances in Health Sciences Education, 10*, 253-261.

Dweck, C.S. (2000). *Self-theories: Their role in motivation, personality, and development.* Philadelphia: Psychology Press.

Entwistle, N. (1995). Frameworks for understanding as experienced in essay writing and in preparing for examinations. *Educational Psychology*, *30*, 47-54.

Fay, A.L., & Mayer, R.E. (1994). Benefits of teaching design skills before teaching LOGO computer programming: Evidence for syntax independent learning. *Journal of Educational Computing Research, 13*, 187-210.

Feltovich, P.J., Spiro, R.J., & Coulson, R.L. (1993). Learning, teaching, and testing for complex conceptual understanding. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests.* Hillsdale: Lawrence Erlbaum Associates, Publishers.

Ferrall, C. (1995). Interactive statistic tutorials in stata. *Journal of Statistics Education, 3*.

Friedman, H. H., Friedman, L. W., & Amoo, T. (2002). Using humor in the introductory statistics course. *Journal of Statistics Education, 10*.

Gal, I., & Garfield, J. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal and J. Garfield (Eds.), *The assessment challenge in statistics education*. Amsterdam: IOS Press.

Gal, I., Ginsberg, L., & Schau, C. (1997). Monitoring attitudes and believes in statistics education. In I. Gal and J. Garfield (Eds.), *The Assessment Challenge in Statistics Education*, Amsterdam: IOS Press.

Garfield, J. B. (1993). Teaching statistics using small group cooperative learning. *Journal of Statistics Education, 1*.

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2(1)*, 22-38.

Garfield, J. B., & Chance, B. L. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning, 2*, 99-125.

Garfield, J. B., delMas, R.C., Chance, B. L. (1999). *The role of assessment in research on teaching and learning statistics*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Gerjets, P., Scheiter, K., & Catrambone, R. (2006). Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations? *Learning and Instruction, 16*, 104-121.

Gimino, A.E., (2002, April). Students' investment of mental effort. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Gijbels, D., Dochy, F., van den Bossche, P., & Segers, M. (2005). Effects of problem-based learning: a meta-analysis from the angle of assessment. *Review of Educational Research, 75*, 27-61.

Giraud, G. (1997). Cooperative Learning and Statistics Instruction. *Journal of Statistics Education, 5*.

Glaser, R. (1991). The maturing of the relationship between the science of learning and cognition and educational practice. *Learning and Instruction, 1*, 129-144.

Glaser, R. (1993). *Advances in Instructional Psychology.* Hillsdale: Lawrence Erlbaum Associates

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

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetition on recall and recognition. *Memory & Cognition, 7*, 95-112.

Glenberg, A. M., Sanocki, Th., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology, 116,* 119-136.

Goldsmith, T. E., Johnson, P. J. (1990). A structural assessment of classroom learning. In R., W. Schvaneveldt (Ed), *Pathfinder associative networks. Studies in knowledge organization* (p. 1-30). Norwood, Ablex Publishing Corporation.

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing structural knowledge. *Journal of Educational Psychology*, *83*, 88-96.

Graesser, A. C., Bagget, W., Williams, K. (1996). Question driven explanatory reasoning. *Applied Cognitive Psychology, 10*, S17-S131.

Graesser, A. C., Olde, B., & Lu, S. (2001). *Question-driven explanatory reasoning about devices that malfunction.* Paper presented at the Annual Meeting of the American Educational Research Association 2001. Seattle, April 10-14, 2001

Graesser, A. C., Person, N. K., Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*, 495-522.

Graham, S., & Weiner, B. (1987). Some educational implications of sympathy and anger from an attributional perspective. In Snow, R.E., Farr, M.J. (Eds), *Aptitude, learning, and instruction. Volume 3: Conative and affective process analyses* (pp. 199-221). Hillsdale: Lawrence Erlbaum Associates, Inc., Publishers.

Greeno, J.G. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53*, 5-26.

Hmelo, C. E., & Evensen, D. H. (2000). Introduction. Problem-based learning: Gaining insights on learning interactions through multiple methods of inquiry. In D. H. Evensen and C. E. Hmelo (Eds.), *Problem-Based Learning: A research perspective on learning interactions*. Mahwah: Lawrence Erlbaum Associates Publishers.

Holcomb, J., & Spalsbury, A. (2005). Teaching students to use summary statistics and graphics to clean and analyze data. *Journal of Statistics Education, 13*.

Hu, L., Bentler,P., M., & Kano,Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological-Bulletin, 112*, 351-362.

Isen, A.M., Daubman, K.A., Gorgoglione, J.M. (1987). The influence of positive affect on cognitive organisation: implications for education. In Snow, R.E.,

Farr, M.J. (Eds), *Aptitude, learning, and instruction. Volume 3: Conative and affective process analyses* (pp. 143-164). Hillsdale: Lawrence Erlbaum Associates, Inc., Publishers.

Johnson, H. D., & Dasgupta, N. (2005). Traditional versus non-traditional teaching: perspectives of students in introductory statistics classes. *Journal of Statistics Education, 13*.

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge.* Hillsdale: Lawrence Erlbaum Associates, Publishers.

Jöreskog, K.G., & Sörbom, D. (1989). *LISREL 7. A guide to the program and applications.* Chicago: Scientific Software, Inc.

Kahn, M. (2005). An exhalent problem for teaching statistics. *Journal of Statistics Education, 13*.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93*, 579-588.

Keeler, C. M., & Steinhorst, P. K. (1995). Using Small Groups to Promote Active Learning in the Introductory Statistics Course: A Report from the Field. *Journal of Statistics Education, 3*.

Kelly, G.A. (1958). Man's construction of his alternatives. In G. Lindzey (Ed.), *Assessment of human motives.* New York: Grove.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological review, 9*, 163-182.

Kintsch, W. (1998). *Comprehension. A paradigm for cognition.* Cambridge: Cambridge University Press.

Kintsch, W. (1998b). The representation of knowledge in minds and machines. *International Journal of Psychology*, *33*, 411-420.

Kirschner, P.A., Sweller, J. & Clark, R.E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist discovery, problem-based, experiential and inquiry-based teaching. *Educational Psychologist, 41*, 75-86.

Kraiger, K., & Cannon-Bowers, J. A. (1995). Measuring knowledge organization as a method for assessing learning during training. *Human factors, 37*, 804-816.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*, 115- 129.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem : The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104,* 211-240.

Larsen, M. D. (2006). Advice for new and student lecturers on probability and statistics. *Journal of Statistics Education, 14*.

Lee, M., & Thompson, A. (1997). Guided instruction in LOGO programming and the development of cognitive monitoring strategies among college students. *Journal of Educational Computing Research, 16*, 125-144.

Lepper, M. R., Drake, M. F., & O'Donnell-Johnson, T. (1997).Scaffolding techniques of expert human tutors. In K. Hogan and M. Pressley (Eds.),

*Scaffolding student learning: instructional approaches and issues*. Cambridge: Brookline Books.

Magel, R. C. (1998). Using cooperative learning in a large introductory statistics class. *Journal of Statistics Education, 6*.

Malone, T.W., & Lepper, M.R. (1987). Making learning fun: a taxonomy of intrinsic motivations for learning. In Snow, R.E., Farr, M.J. (Eds), *Aptitude, learning, and instruction. Volume 3: Conative and affective process analyses* (pp. 223-253). Hillsdale: Lawrence Erlbaum Associates, Inc., Publishers.

Mannes, S. M., & Kintsch, W. (1987). Knowledge organization and text Organization. *Cognition and Instruction*, *4*, 91-115.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Mayer, R., E. (1989). Models for understanding. *Review of Educational Research*, *59*, 43-64.

Mayer, R. E. (1992). Knowledge and thought: Mental models that support scientific reasoning. In R. A. Duschl, & R. J. Hamilton (Eds.), *Philosophy of science, cognitive psychology, and educational theory and practice*. Albany, State University of New York Press.

Mayer, R. E. (1997). *Thinking, problem solving, cognition*. New York: W. H. Freeman and Company.

Mayer, R.E. (2004). Should there be a three-strike rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist, 59*, 14-19.

Mayer, R. E., & Wittrock, M. C. (1996). Problem solving transfer. In D. C. Berliner, & R. C. Calfee (Eds.), *Handbook of educational psychology*. New York: Prentice Hall.

McNamara, D. S., Kintsch, E., Butler-Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*, 1-43.

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior, 9*, 596-606.

Merrill, D. C., Reiser, B. J. Merrill, S. K., & Landes, S. (1995). Tutoring: Guided learning by doing. *Cognition and Instruction, 13*, 315-372.

Merrill, D.C., Reiser, B.J., Ranney, M., & Trafton, J.G. (1992). Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *The Journal of the Learning Sciences, 2*, 277-305.

Murphy, P. K., & Alexander, P. A. (2000). A motivated exploration of motivation terminology. *Contemporary Educational Psychology, 25*, 3-53.

Mvududu, N. (2003). A cross-cultural study of the connection between students' attitudes toward statistics and the use of constructivist strategies in the course. *Journal of Statistics Education, 11*.

Neisser, U. (1984). Interpreting Harry Bahrick's Discovery: What confers immunity against forgetting. *Journal of Experimental Psychology, 113*, 32-35.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

Noordman, L., G., M., & Vonk, W. (1998). Memory-based processing in understanding causal information. *Discourse Processes*, *26*, 191-212.

Norman, (1982). *Learning and memory.* San Francisco: W. H. Freeman.

Norman, D. A., Gentner, S., & Stevens, A. L. (1976). Comments on learning schemata and memory representation. In D. Klahr (Ed.). *Cognition and Instruction.* Hillsdale: Lawrence Erlbaum Associates, Publishers.

Novak, J. D., (2002). Meaningful learning: the essential factor for conceptual change in limited or inappropriate propositional hierarchies leading to empowerment of learners. *Science Education, 86*, 548-571.

Olson, J., & Biolsi, K., J. (1991). Techniques for representing expert knowledge. In K. A. Ericsson & J. Smith (Eds.). *Toward a general theory of expertise.* Cambridge, UK, Cambridge University Press.

Paas, F.G.W.C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84,* 429-434.

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional Science, 32*, 1-8.

Paas, F., Tuovinen, J.E., Tabbers, H., & Van Gerven, P.W.M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Pychologist, 38*, 63-71.

Paas, F.G.W.C., & van Merriënboer, J.J.G. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors, 35*, 737-743.

Paas, F., & van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive load approach. *Journal of Educational Psychology, 86*, 122-133.

Paas, F.G.W.C., & van Merriënboer, J.J.G., Adam, J.J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills, 79*, 419-430.

Peterson, C., Maier, S.F., & Seligman, M.E.P. (1993). *Learned helplessness. A theory for the age of personal control.* New York: Oxford University Press.

Phye, G. D., (1992). Strategic transfer: A tool for academic problem solving. *Educational Psychology Review, 4*, 393-421.

Phye, G. D. (1997). *Handbook of academic learning : construction of knowledge*. San Diego: Academic Press.

Pintrich, P.R. (2000). The roal of goal orientation in self-regulated learning. In M. Boekaerts, P.R. Pintrich, & M. Zeidner (Eds), *Handbook of self regulation* (pp. 451-502). San Diego: Academic Press.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95*, 667-686.

Pintrich, P.R., & Schunk, D.H. (1996). *Motivation in education: theory, research and applications.* Upper Saddle River, NY: Merrill, Prentice Hall.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.). *Semantic information processing.* Cambridge: MIT Press.

Reimann, P., & Schult, T. J. (1996). Turning examples into cases: Acquiring knowledge structures for analogical problem solving, *Educational Psychologist, 31*, 123-132.

Renkl, A. (1997). Learning from worked-out examples: a study on individual differences. *Cognitive Science, 21*, 1-29.

Renkl, A. (1999). Learning mathematics from worked-out examples: Analyzing and fostering self-explanations. *European Journal of Psychology and Education, 14*, 477-488.

Renkl, A. (2002). Worked-out examples: Instructional explanations support learning by self-explanations. *Learning and Instruction, 12*, 529-556.

Renkl, A., Mandl, H., & Gruber, H. (1996). Inert knowledge: analysis and remedies. *Educational Psychologist, 31*, 115-121.

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary Educational Psychology, 23*, 90-108.

Rowe, A. L., Cooke, N. J., Hall, E. P., & Halgren, T. L. (1996). Toward an on-line knowledge assessment methodology: Building on the relationship between knowing and doing. *Journal of Experimental Psychology: Applied*, *2*, 31-47.

Ruiz-Primo, M. A., Schultz, S. E., & Shavelson, R. J. (2001). Comparison of the reliability and validity of scores from two concept mapping techniques. *Journal of Research in Science Education, 38*, 260-278.

Rumelhart, D. E., & Norman, D. A. (1983). Representation in memory. In R. C. Atkinson, R.J. Herrnstein, & R. D. Luce (Eds.). *Handbook of experimental psychology.* New York: Wiley.

Rumsey, D.J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, 10(3).

Satorra, A., & Bentler, P.M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C.C. Clogg (Eds), *Latent variables analysis: Application for developmental research* (pp. 399-419). Thousand Oaks: Sage.

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals, and understanding.* Hillsdale: Lawrence Erlbaum Associates, Publishers.

Schau, C., & Mattern, N. (1997). Assessing students' connected understanding of statistical relationships. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education*. Amsterdam: IOS Press.

Schau, C., Stevens, J., Dauphinee, T.L., & Del Vecchio, A. (1995). The development and validation of the Survey of Attitudes Toward Statistics. *Educational and Psychological Measurement,55,* 868-875.

Schmidt, H. G., v.d. Arend, A., Moust, J. H. C., Kokx, I., & Boon, L. (1993). Influence of tutors' subject-matter expertise on student effort and achievement in Problem-based Learning. *Academic Medicine, 68*, 784-791.

Schmidt, H. G., & Boshuizen, H. P. A. (1993). On acquiring expertise in medicine. *Educational Psychology Review, 5*, 205-221.
Schmidt, H. G., & Moust, J. H. C. (1995). What makes a tutor effective? A structural-equations modelling approach to learning in problem-based curricula. *Academic Medicine, 70*, 708-714.
Schmidt, H. G., & Moust, J. H. C. (2000). Factors affecting small-group tutorial learning: A review of research. In D. H. Evensen and C. E. Hmelo (Eds.), *Problem-Based Learning: A research perspective on learning interactions*. Mahwah: Lawrence Erlbaum Associates Publishers.
Schvaneveldt, R., W. (1990). *Pathfinder associative networks. Studies in knowledge organization*. Norwood, Ablex Publishing Corporation.
Schwartz, N. (1999). Self-reports. How the questions shape the answers. *American Psychologist, 54*, 93-105.
Scott Long, J. (1983). *Covariance structure models.* Newbury Park: Sage Publications Inc.
Seabrook, R., Brown, G. D. A., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Applied Cognitive Psychology, 19*, 107-122.
Seegers, G., & Boekaerts, M. (1993). Task motivation and mathematics achievement in actual task situations. *Learning and Instruction, 3*, 133-150.
Semb, G. B., & Ellis, J. A. (1994). Knowledge taught in school: What is remembered? *Review of Educational Research, 64*, 253-286.
Snijders, T. A. B., & Bosker, R. J. (2003). *Multilevel Analysis. An introduction to basic and advanced multilevel modelling*. London: SAGE Publications Ltd.
Steffe, L.P., & Gale, J. (1995). *Constructivism in education.* Hillsdale: Erlbaum.
Steinhorst, R. K., & Keeler, C. M. (1995). Developing material for introductory statistics courses from a conceptual active learning viewpoint. *Journal of Statistics Education, 3*.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257285.
Sweller, J. (1999). *Instructional design in technical areas.* Mahwah, NJ: Erlbaum.
Sweller, J., & Cooper, G.A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2,* 59-89.
Sweller, J., van Merriënboer, J.J.G., & Paas, F.G.W.C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-296.
Tuovinen, J.E. & Paas, F. (2004). Exploring multidimensional approaches to the efficiency of instructional conditions. *Instructional Science, 32*, 133-152.
Tuovinen, J.E., Sweller, J. (1999). A comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology, 91,* 334-341.
Valås, H., & Søvik, N. (1993). Variables affecting students' intrinsic motivation for school mathematics: two empirical studies based on Deci and Ryan's theory on motivation. *Learning and Instruction, 3*, 281-298.

Van de Wiel, M. W. J., Boshuizen, H., P., A., & Schmidt, H., G. (2000). Knowledge restructuring in expertise development: Evidence from pathophysiological representations of clinical cases by students and physicians. *European Journal of Cognitive Psychology*, *13*, 323-355.

van Gog, T., Paas, F., & van Merriënboer, J.J.G. (2006). Process-oriented worked examples: Improving transfer performance through enhanced understanding. *Instructional Science, 32*, 83-98.

Vanhoof, S., Castro Sotos, A. E., Onghena, P., Verschaffel, L., & Van Dooren, W. (2006). Attitudes toward statistics and their relation with short- and long-term exam results. *Journal of Statistics Education, 14*.

van Merriënboer, J.J.G., & Sweller, J. (2005). Cognitive load theory and complex learning: recent developments and future directions. *Educational Psychology Review, 17*, 147-177.

Vermetten, Y., J., Vermunt, J., D., & Lodewijks, H., G. (2002). Powerful learning environments? How University students differ in their response to instructional measures. *Learning and Instruction, 12*, 263-284.

Vermunt, J. (1994). Design principles of process-oriented instruction. In F. P. C. M. de Jong & B. H. A. M. van Hout-Wolters (Eds.), *Process-oriented instruction and learning from text*. Amsterdam: VU University Press.

Vermunt, J. D., & Vermetten, Y. J. (2004). Patterns in student learning: Relationships between learning strategies, conceptions of learning, and learning orientations. *Educational Psychology Review, 16*, 359-384.

Volet, S. E. (1991). Modelling and coaching of relevant metacognitive strategies for enhancing university students' learning. *Learning and Instruction, 1*, 319-336.

Volet, S.E. (1997). Cognitive and affective variables in academic learning: the significance of direction and effort in students' goals. *Learning and Instruction, 7*, 235-254.

Volet, S., McGill, T., & Pears, H. (1995). Implementing process-based instruction in regular university teaching: Conceptual, methodological, and practical issues. *European Journal of Psychology of Education, 10*, 385-400.

von Hippel, P. T. (2005). Mean, median, and skew: correcting a textbook rule. *Journal of Statistics Education, 13*.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: a study of conceptual change in childhood. *Cognitive Psychology, 24,* 535-585.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge: Harvard University press.

Walters, E. J., Morrell, C. H., & Auer, R. E. (2006). An investigation of the median-median method of linear regression. *Journal of Statistics Education, 14*.

Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction, 7*, 1-39.

Watkins, D. (1996). The influence of social desirability on learning process questionnaires: A neglected possibility? *Contemporary Educational Psychology, 21*, 80-82.

Weiner, B. (1986). *An attributional theory of motivation and emotion.* New York: Springer Verlag.

Weiner, B. (1992). *Human motivation. Metaphors theories, and research.* Newbury Park: Sage Publications Inc.

West, R. W., & Ogden, R. T. (1998). Interactive demonstrations for statistics education on the world wide web. *Journal of Statistics Education, 6*.

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

Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and text by latent semantic analysis. *Discourse Processes*, *25*, 309-336.

Wyman, B. G., & Randel, J. M. (1998). The relation of knowledge organization to performance of a cognitive task. *Applied Cognitive Psychology*, *12*, 251-264.

## Samenvatting

Het doel van de studies in deze dissertatie was om te onderzoeken welke interventies in het bestaande statistiekonderwijs zouden leiden tot een beter inzicht van de studenten in de materie.
Deze doelstelling vereiste ten eerste een goede afbakening, en definiëring van de term inzicht (conceptual understanding). Daarnaast vereiste dit een meetinstrument dat juist die eigenschappen van kennis in kaart brengt, die volgens de gehanteerde definitie aantonen dat er sprake is van inzicht. In **hoofdstuk 2** wordt deze kwestie besproken. De verschillende stappen, die achtereenvolgens genomen moeten worden om inzicht te meten, zijn:
1. Afbakening en definiëring van de terminologie.
2. Bepalen van criteria die gebaseerd zijn op de gehanteerde definities.
3. Het kiezen van een geschikte methode voor het externaliseren van de kennis van de studenten. Dat wil zeggen dat met deze methode de eigenschappen van de kennis in kaart te brengen zijn, waar de vastgelegde criteria betrekking op hebben.
4. Een methode om te beoordelen in welke mate de geëxternaliseerde kennis aan de criteria voldoet.
5. Een geschikt scoringssysteem.

Aan de hand van een voorbeeld en een concrete toepassing wordt deze benadering besproken. De conclusie is dat de voorgestelde methode redelijk betrouwbare en valide metingen oplevert. In een aantal studies van deze dissertatie is deze methode gebruikt.
In **hoofdstuk 3** is aan de hand van een tweetal motivationele theorieën een model opgesteld, dat de relatie tussen de wijze waarop studenten hun falen/slagen verklaren, de uitkomstverwachtingen, het studiegedrag, een aantal affectieve aspecten en hun presteren, weergeeft. Dit model is vervolgens, aan de hand van verzamelde data, getoetst door middel van een zogenaamde padanalyse (Lisrel). Uit de resultaten blijkt dat affect een centrale rol speelt. Wanneer studenten statistiek, een interessant, relevant en aardig vak vinden, dan presteren ze beter. Affectieve aspecten blijken vooral beïnvloed te worden door het idee dat er stabiele oorzaken zijn voor het falen. Wanneer de student denkt geen controle te hebben over zijn presteren, dan zal hij weinig effect verwachten van zijn inspanningen. Het blijkt bovendien dat de prestaties van een student(e) zullen verbeteren wanneer hij/zij persisteert, dat wil zeggen niet te gauw het hoofd in de schoot legt,. Voor de praktijk betekent dit dat wanneer studenten succeservaringen opdoen en gestimuleerd worden door te zetten, het presteren indirect en direct verbeterd kan worden.
In **hoofdstuk 4** is het effect van een meer directieve rol van de tutor op een tweetal motivationele aspecten en het presteren van de studenten bestudeerd. Tijdens de onderwijsbijeenkomsten in het probleem gestuurd onderwijs (PGO) van statistiek werd aan tutoren gevraagd sturende vragen te stellen om studenten te stimuleren zich te concentreren op de belangrijkste onderwerpen en om de discussies meer te structureren. De resultaten tonen aan dat de motivatie

duidelijk verbeterde. Het presteren, als gemeten door middel van het behaalde resultaat op de toets, verbeterde marginaal.
In **hoofdstuk 5** zijn twee interventies getest: het effect van het spreiden van studieactiviteiten (distribution of practice) en dezelfde directieve rol van de tutor als in hoofdstuk 4. Als uitkomstmaat is nu gebruikt het inzicht van studenten zoals gemeten met behulp van de in hoofdstuk 2 voorgestelde methode. Uit de resultaten blijkt dat wanneer studenten meer in de gelegenheid zijn om hun studieactiviteiten te spreiden, dat hun inzicht in de materie groter is en dat de retentie van het geleerde groter is. Ditzelfde geldt voor een meer directieve rol van de tutor; meer inzicht en een betere retentie. Hierbij dient overigens opgemerkt te worden dat de retentie betrekkelijk gering is. Wat studenten 6 maanden na de introductiecursus nog weten van de behandelde stof, is slechts marginaal meer dan aan het begin van de cursus.
In **hoofdstuk 6** is een experiment beschreven. De studies in de voorgaande hoofdstukken zijn uitgevoerd als veldstudies in het bestaande onderwijs. In dit experiment is het effect onderzocht van geschreven directieve vragen als zijnde een gestandaardiseerde vorm van sturing. Concreet werd onderzocht of deze vragen via de mediërende variabelen mentale belasting (cognitive load) en zelfuitleg (self-explanations) zou leiden tot meer inzicht en vervolgens tot een positievere attitude (hier gepostuleerd als motivationele component). Uit het resultaat bleek dat de directieve vragen inderdaad leiden tot meer inzicht. Bovendien kon de mediërende rol van mentale belasting en zelfuitleg bevestigd worden. Hiermee is niet alleen het resultaat van de veldstudies ondersteund, er is tevens een mogelijk werkingsmechanisme aangetoond. Het verwachtte effect van de directieve vragen op attitude werd niet gevonden. Wel werd er een significante correlatie gevonden tussen meer inzicht en het presteren op de toets van een voorgaande cursus enerzijds en een positieve attitude anderzijds.
Op basis van de beschreven resultaten kan gezegd worden dat directieve vragen een positief effect hebben op inzicht en mogelijk op de attitude van studenten. Aangezien attitude sterk affectieve componenten heeft en affect een belangrijk gegeven is voor het presteren van studenten in het statistiekonderwijs, zijn directieve vragen een effectief middel om inzicht en presteren zowel direct als indirect te verbeteren.

## Dankwoord

Graag wil ik iedereen bedanken die op enigerlei wijze heeft bijgedragen aan de totstandkoming van dit proefschrift. Oprecht en welgemeend bedankt iedereen.

Met name een woord van dank aan mijn promotor en begeleiders. Martijn bedankt voor de aanwijzingen die de diverse artikelen meer publicabel gemaakt hebben en je adviezen ten aanzien van het afsluitende experiment. Margje bedankt voor de minutieuze feedback, die je veel inzet en tijd gekost hebben en je gedrevenheid tijdens de overlegsituaties. Tjaart bedankt, zonder jou was dit proefschrift er nooit gekomen en had ik nu niet op deze plek gezeten.

De mooie omslag van mijn proefschrift is verzorgd door mijn vriend Frenk, bijzonder woord van dank hiervoor.

Mijn collega's van Methodologie en Statistiek bedankt voor de aangename en inspirerende omgeving, waarin ik aan dit proefschrift mocht werken. Eigenlijk wilde ik geen namen noemen, maar vooruit; Nick toch bedankt.
Mijn collega's van Onderwijsontwikkeling en –research bedankt voor de enthousiaste discussies tijdens onze lunchbijeenkomsten. En natuurlijk bedankt voor de financiële bijdragen, die mij in staat stelden de afgelopen jaren de vele proefpersonen van cadeaubonnen te voorzien.

Mijn vrienden hebben, mogelijk zonder dat ze dat ze zich dat zelf realiseerden, mij ook op diverse manieren gesteund. Mijn medepromovendi omdat ik van hen hoorde dat ook zij geconfronteerd werden met moeilijke fasen in hun promotie, waar ze zich met succes doorheen geworsteld hebben. Dat gaf me nieuwe moed. Mijn medesporters omdat ze het mogelijk gemaakt hebben, me uit te leven in mijn sport en daarna weer fris aan de slag te kunnen gaan. Hein, ik heb genoten van de gesprekken die ik met je gevoerd heb. Mijn vrienden en vriendinnen waar ik veel om geef. Allemaal bedankt. Ik wil jullie graag en veel blijven zien.

Gwen, jij bent in mijn leven gekomen op het moment dat ik in de afrondende fase zat van dit proefschrift; de spreekwoordelijke laatste loodjes. Je was betrokken en je hebt me gemotiveerd en gesteund. Dankjewel. Je hebt er begrip voor getoond dat ik weer eens weinig tijd had of gewoon wat afwezig was. Ik hoop dat we er nu meer voor elkaar kunnen zijn. Ik vind het heerlijk dat je bij me bent. Ik hou van jou.

Pa en Ma, het was een hele ommezwaai, maar jullie hebben steeds vertrouwen in mij uitgestraald. Bedankt.
Anne en Job, bedankt voor wie jullie zijn. Ik ben er trots op dat ik jullie vader ben.

Luc Budé, Maastricht, juni 2007

## Curriculum vitae

Luc Budé werd geboren op 18 juli 1959 te Beek. Na het behalen van zijn VWO diploma in 1978 heeft hij aansluitend fysiotherapie gestudeerd te Heerlen. Gedurende een periode van 15 jaar heeft hij gewerkt als fysiotherapeut in zijn eigen praktijk. In 1997 is hij cognitieve psychologie gaan studeren aan de Universiteit Maastricht. Hij behaalde hiervoor in 2001 cum laude zijn diploma betreffende de beide afstudeertrajecten cognitieve ergonomie en onderwijspsychologie. De afsluitende onderzoeksstage was gericht op het meten van begrip en werd begeleid door Dr. M. W. J. van de Wiel en prof. H. G. Schmidt. In 2001 werd hij aangesteld als promovendus bij de capaciteitsgroep Methodologie en Statistiek. Het onderzoek naar statistiekonderwijs dat hij hier verrichtte, resulteerde in dit proefschrift. Sinds 1 juni 2007 is hij aangesteld als docent bij dezelfde capaciteitsgroep.