

**Propositional knowledge for
conceptual understanding of statistics**

This research was carried out at



In the School of Health Professions Education



© Jimmie Leppink, Maastricht 2012

Production: Boekenplan | Maastricht

ISBN 978 90 8666 254 8



Propositional knowledge for conceptual understanding of statistics

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 woensdag 20 juni om 12.00 uur

door

Jimmie Leppink

Promotores

Prof. dr. M. P. F. Berger

Prof. dr. C. P. M. Van der Vleuten

Copromotor

Dr. N. J. Broers

Beoordelingscommissie

Prof. dr. D. H. J. M. Dolmans (voorzitter)

Dr. A. Bakker, Statistics Education, Freudenthal Instituut, Universiteit Utrecht

Prof. dr. W. H. Gijsselaers

Prof. dr. G. W. C. Paas, Onderwijspsychologie i.h.b. Cognitie van het Leren, Erasmus Universiteit
Rotterdam

Dr. D. T. Tempelaar

Table of contents

Chapter 1: Introduction <i>Based on paper for the 8th International Conference on Teaching Statistics (2010)</i>	6
Chapter 2: Task-and student-related factors in propositional manipulation <i>Journal of Statistics Education (2011)</i>	16
Chapter 3: The effect of guiding students into self-explanation <i>Submitted</i>	31
Chapter 4: The expertise reversal effect (I): self-explanation <i>Higher Education (2011)</i>	52
Chapter 5: The expertise reversal effect (II): joint explanation <i>Educational Research and Evaluation (2012)</i>	66
Chapter 6: Propositional manipulation in a statistics lecture <i>Submitted</i>	80
Chapter 7: Guided problem-based learning of statistics <i>Submitted</i>	92
Chapter 8: Discussion	108
English summary	116
Nederlandse samenvatting	118
References	121
Appendix	126
Acknowledgements	128
About the author	130
SHE Dissertation Series	131

Chapter 1

Introduction

Based on

Leppink, J. (2010). Adjusting cognitive load to the student's level of expertise for increasing motivation to learn, *Proceedings of the Eighth International Conference on Teaching Statistics*, Ljubljana, Slovenia [open online access]

1.1. Statistics as a subject in empirical disciplines

Across a wide range of university disciplines, statistics is considered to be an indispensable part of the curriculum. Whether students follow training to become an expert in biochemistry, business administration, psychology, medicine, or political science, at some point or another, they are likely to encounter one or more mandatory courses in statistics in their studies.

A combination of factors contributes to the finding that many students develop only superficial understanding of this subject. To begin with, in many curricula the subject is given very limited time (Van Buuren, 2008), and part of that time is used to make students familiar with statistical software (Hulsizer & Woolf, 2009). Given that the domain of statistics is a complex knowledge domain that is characterized by abstract and hierarchical concepts which not always have a direct or simple meaning outside the domain, avoiding or surfacing the subject matter can easily lead to disorientation. Moreover, since for a proper understanding of the subject matter students are required to understand a number of formulae and mathematical relationships, especially students with a non-mathematical background (e.g., students in the social sciences) tend to avoid the subject matter (Broers, 2009). Adjusted teaching methods are needed to help these students develop knowledge and understanding of statistics from the very start and step by step.

1.2. From propositional knowledge to conceptual understanding

A study by Huberty, Dresden, and Bak (1993) revealed three dimensions underlying statistical knowledge: computational aptitude, propositional knowledge, and conceptual understanding. Computational aptitude includes the ability to understand formulas and use them correctly. Propositional knowledge refers to more or less isolated knowledge of statistical concepts, propositions, and ideas related to statistics. Conceptual understanding is the ability of students to perceive links and interrelationships between the various concepts, propositions, and ideas.

In the past few decades, statistics education has profited from the revolution in teaching that was due to the advent of the personal computer. Whereas in former days statistics education largely focused on teaching students to do calculations by hand, the computational power of modern statistical software has reduced traditional number crunching. Nowadays, students work on large and realistic datasets, with the emphasis on the interpretation of the statistical output. Students are taught to make intelligent use of the multitude of quantitative information and learn to engage in statistical reasoning. For this reason, the focus of statistics education shifted from computational aptitude to conceptual understanding.

When studying the statistical literature, attending a lecture, or when performing a learning task on statistics, students are confronted with important concepts and core ideas. Students first have to isolate the important ideas by deriving and studying their constituent elements (i.e., propositional knowledge), and then relate and integrate these elements into schemata (i.e., conceptual understanding) and gradually develop an integrated knowledge network (Novak, 2002). Propositional knowledge appears to be a necessary, but not a sufficient condition for obtaining conceptual understanding (Marshall, 1995). Further knowledge elaboration is needed to help students develop conceptual understanding.

Following Bude (2007), the focus in this thesis is on the improvement of students' conceptual understanding of statistics. The definition for conceptual understanding in both theses is the same. However, motivational and assessment aspects play a more prominent role in the thesis by Bude than in the current thesis. The current thesis focuses on different instructional approaches to help students develop propositional knowledge and conceptual understanding.

1.3. Conceptual understanding through knowledge elaboration at different levels

Knowledge elaboration involves using prior knowledge to structure and integrate new information. Self-explanation (Atkinson, Renkl, & Merrill, 2003; Berthold & Renkl, 2009) and argumentation (Fischer, 2002; Knipfer, Mayr, Zahn, Schwan, & Hesse, 2009) are well-known knowledge elaboration processes. The two most successful cognitive mechanisms of self-explanations appear to be filling knowledge gaps (Chi, 2000) and constructing knowledge networks (Novak, 2002). Further, research suggests that explanation to peers leads to more effective learning than self-explanation (Kramarski & Dudai, 2009), although other studies find no differences (Moreno, 2009) or find that self-explanation leads to deeper learning than explanation to peers (Hausmann, Van de Sande, & Van Lehn, 2008). In line with the aforementioned findings, a new method has been proposed: the method of propositional manipulation (MPM).

MPM is based on the reasoning that self-explanation and argumentation can help students to develop conceptual understanding of a complex knowledge domain. Although MPM has been developed for the statistics knowledge domain and tested empirically within this domain this format may be applicable in other complex knowledge domains as well.

MPM comprises three steps. In the first step, the instructor determines the subject matter and divides it into a limited number of propositions. Propositions are statements referring to single statistical ideas and concepts (e.g., arithmetic mean, z-score, sampling distribution, expected value). The number of propositions needed depends on size and content of the subject matter. Further, the exact content and formulation of the propositions should depend on students' statistics proficiency level. Teaching to students at a more elementary level requires more elementary propositions than teaching to students at a more advanced level. The instructor then formulates for each proposition one open-ended question, in such a way that the correct proposition would be the answer to the question.

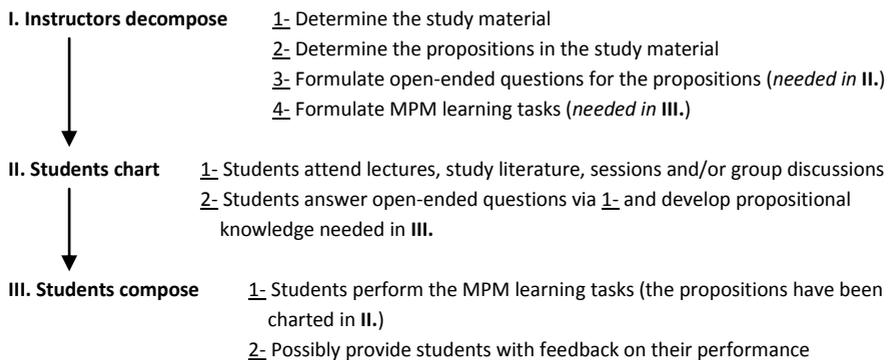
In the second step, students are instructed to answer the open-ended questions formulated by the instructor. Students are provided with the questions (e.g., 'what is a z-score?', 'what is a sampling distribution?', 'what is the expected value of an estimator?'), and they can find the actual propositions in the literature or in lectures on the subject matter (e.g., 'a z-score is the deviation of a score from the arithmetic mean, relative to the standard deviation'). To have

students answer questions referring to propositions is supposed to make them aware of important misconceptions and to help them develop the propositional knowledge that is needed to develop conceptual understanding. Students are stimulated to self-explain the subject matter and they are guided into this process of self-explanation by means of the questions. Given the abstract and cumulative nature of statistics and the frequent and persistent misconceptions students have about the subject matter, this second step of MPM is a necessary step towards developing conceptual understanding of statistics (Broers, 2002).

It is only in the third step of MPM that students begin to develop conceptual understanding, namely by performing a series of MPM learning tasks. In an MPM learning task, students have to relate and integrate a number of propositions into an argument that proves a given hypothesis to be either true or false. In contrast to propositions, the hypothesis typically comprises multiple statistical ideas and concepts. Therefore, hypotheses are generally of higher complexity level than propositions. Once students master the propositions (i.e., propositional knowledge) they should relate and integrate these propositions into an argument in such a way that the argument enables them to understand why the hypothesis is true or false (i.e., conceptual understanding).

Box 1.1.

The method of propositional manipulation (MPM): decompose, chart, and compose



The propositions have been determined by the instructor in the first step. For each proposition, the instructor formulates one question. In the second step, students answer these questions and thereby discover the propositions. In the third step, the instructor gives the students a hypothesis and attaches several propositions formed as questions to this hypothesis. The complexity level and the exact formulation of the hypothesis depend on the learning goals on the statistics course: what interrelationships between statistical ideas and concepts do we want students to know at the end of the course? Next, which questions are attached to the hypothesis depends on the learning goals of the statistics course as well as on the specific content of the hypothesis.

Box 1.2.

Example of an MPM learning task

Hypothesis: *Statistical non-significance means that there is no effect.*

- [1] When is a result statistically significant?
 - [2] Why is the p -value a conditional probability?
 - [3] What is a Type I error?
 - [4] What is a Type II error?
 - [5] How are sample size and Type II error probability related?
-

The learning task in Box 1.2. can be presented in two different ways. One way would be to provide students with the hypothesis, a true/false statement, with the instruction to reason why the statement in question is true or false. Another way, in line with the MPM format, is to instruct students to first answer the five open-ended questions to then formulate an argument comprising their answers that proves whether the statement is true or false. All propositions should be part of the argument, even if the hypothesis can be proven correct or incorrect with a subset of these propositions. It is to be expected that the propositions can be useful for more than one reason. First of all, instructing students to evaluate a statement through the analysis of propositions is that these propositions increase the information-richness of students' arguments. Besides, it could help students structure their thoughts and avoid disorientation in the myriad of propositions, concepts, and ideas. Since students who are guided into self-explanation by the propositions do not need to search for the relevant elements in the subject matter themselves, they have more time and capacity available for self-explanation and argumentation. Further, including the propositions stimulates students to self-explain at a more basic level (i.e., the level of propositions) before engaging in self-explanation at the more complex level (i.e., the level of the hypothesis that comprises multiple propositions and links between them). The study of propositions can serve as a checklist for students to the question to what extent they master the propositions that are needed to solve more complex problems.

MPM is supposed to stimulate students to engage in meaningful learning by encouraging them to self-explain the elements underlying the more complex hypothesis. The students must form an argument for the truth or falsity of the hypothesis based on the answers to the questions and the connections between them. Students are not expected to learn the propositions through MPM learning tasks. It is in the previous (i.e., second) step of MPM that the propositions are presented to the students, in lectures as well as in the (course) literature. Besides, questions should be formulated in such a way that they require only short answers and that each question can be related to at least one other question. Given q number of questions, the argument can comprise a maximum of $q(q - 1)/2$ pairwise connections. This means a maximum of three connections in the case of three questions, six connections in the case of four questions, and ten connections in the case of five questions.

The number of valid connections in an argument depends on the exact content of the propositions. Since each proposition refers to a single statistical idea or concept, a correct argument comprises a relevant set of true propositions and can prove a given hypothesis to be

either true or false. Which hypothesis and which questions one chooses for constructing an MPM learning task depends on the learning goals of the statistics course as well as on the students' statistics proficiency level. Thus, when formulating the hypothesis and questions in an MPM learning task, students' statistics proficiency level has to be taken into account. Further, to put a learning task into a (real-life) context, contextual information (e.g., a problem) can precede the hypothesis.

Developing conceptual understanding requires a sound propositional knowledge as well as self-explanation and argumentation, and MPM encompasses all these elements. By having the instructor choose the propositions, students are guided into self-explanation of these propositions, which helps them to develop propositional knowledge. Next, it is the manipulation of propositions in learning tasks that guides students into self-explanation and argumentation on a higher, more complex level, which helps them to develop conceptual understanding.

1.4. Cognitive load

Although self-explanation and argumentation may be effective knowledge elaboration processes, they impose cognitive load on students (Van Merriënboer & Sweller, 2005). Working memory is limited in capacity (Miller, 1956) as well as in duration (Miyake & Shah, 1999). When consciously processing new elements of information, which in complex knowledge domains like statistics are typically interrelated, working memory can be overloaded (Kalyuga, 2009). Cognitive load theory assumes that the available knowledge structures in long-term memory (i.e., prior knowledge) are essential for preventing working memory overload and for guiding cognitive processes when learning (Van Merriënboer & Sweller, 2005). Most people can retain seven plus or minus two chunks of information in their working memory (Miller, 1956). What is considered a chunk of information depends on the students' prior knowledge or available knowledge structures in long-term memory. The size of a chunk is likely to increase as students' prior knowledge increases. Cognitive load imposed on students should therefore be in accordance with their prior knowledge. Cognitive load should never be too high, only what is 'too high' is relative.

Cognitive load consists of three types of load that are assumed to be additive: intrinsic load, germane load, and extraneous load. Intrinsic load depends on task complexity and students' prior knowledge of the subject. This type of load should be manipulated in instructional design by selecting learning tasks that match students' prior knowledge (Kalyuga, 2009). Germane load arises from instructional features that stimulate cognitive processes that are beneficial for learning, whereas all instructional features not directly beneficial for learning impose extraneous load on students. As the intrinsic load imposed on students when studying statistics is usually high, extraneous load should be minimized to avoid cognitive overload (Kalyuga & Hanham, 2010). By minimizing extraneous load and matching intrinsic load to the students' prior knowledge, students can engage in knowledge elaboration processes like self-explanation and argumentation, processes that impose germane load on students.

Although the process of self-explanation is time consuming, if students are given enough time, learning by self-explanation is – in terms of learning outcomes and cognitive load imposed on students – more effective than observational learning, inquiry learning, and hypermedia learning (Eysink, De Jong, Berthold, Kollöffel, Opfermann, & Wouters, 2009). That is, self-explanation can enhance germane load activities more than the other forms of learning. Further, different studies demonstrated that self-explanation enhanced by prompting is more effective

than spontaneous self-explanation (Atkinson et al., 2003; Chi, De Leeuw, Chiu, & LaVancher, 1994). Moreover, it appears that guiding or assisting self-explanation prompts by means of open-ended questions as in MPM is more effective than merely prompting self-explanation (Berthold, Eysink, & Renkl, 2009). An explanation for the latter is that when students are not guided into self-explanation, they need to search the relevant subject matter themselves before they can self-explain, and this can easily lead to disorientation and therefore less efficient learning (Eysink et al., 2009). Since students guided into self-explanation do not need to search for the relevant subject matter themselves, they have more time and capacity available for self-explanation and argumentation.

1.5. The expertise reversal effect

Although self-explanation may be effective for some students, a question that arises is whether novice students have sufficient prior knowledge and argumentation skills to learn from self-explaining the subject matter (Kalyuga, 2009). Students who have insufficient prior knowledge experience an extra high intrinsic load, when confronted with a learning task in which they are guided to elaborate their knowledge (Kalyuga & Hanham, 2010). In this case of high intrinsic load, any additional cognitive activities induced by guiding self-explanation may take cognitive load to the limits of working memory and lead to cognitive overload. From this perspective, it is not surprising that previous studies on learning from worked-out examples indicate that novice students who have insufficient or partly incorrect prior knowledge learn more from studying worked-out examples than from solving problems or imagining solution steps themselves (e.g., Cooper, Tindall-Ford, Chandler, & Sweller, 2001; Kalyuga, Chandler, Tuovinen, & Sweller, 2001b; Lovett, 1992). An explanation for the latter is that for learning tasks with high intrinsic load, problem-solving imposes high extraneous load for novice learners (Paas & Van Gog, 2006; Sweller, Van Merriënboer, & Paas, 1998).

People tend to solve new problems by searching for similar problems – of which the solution is known and the solution steps have been worked-out – that can guide their solution of the new problems (Mayer, 1992). Worked-out examples of problems can guide students into self-explanation, but it depends on the students' prior knowledge (Kalyuga et al., 2001b) as well as on the design of the examples and the instructions in the examples whether students actually learn by doing so (Paas & Van Merriënboer, 1994; Van Merriënboer, Schuurman, De Croock, & Paas, 2002). Thus, considering students' prior knowledge is important, since it influences the effectiveness of ways to increase germane load activities like self-explanation (Paas & Van Gog, 2006). The learning activities that are intended to induce germane load will only do so if they are at a suitable level of difficulty for the student. With more prior knowledge, worked-out examples become redundant and problem-solving becomes superior (Kalyuga et al., 2001b). When a learner is able to self-explain, instructional explanations as provided in worked-out examples impose extraneous load instead of germane load on the students (Kalyuga, Ayres, Chandler, & Sweller, 2003). The latter is also called the expertise reversal effect: there is an interaction between the levels of students' prior knowledge and effectiveness of different instructional methods, meaning that instructional methods that are effective for students who have little prior knowledge can lose their effectiveness and even have negative consequences for more knowledgeable students (Kalyuga, 2005, 2006, 2007; Kalyuga et al., 2003; Kalyuga, Chandler, & Sweller, 2001a).

1.6. Worked-out examples and propositional manipulation

If performing MPM learning tasks is too complex for a group of students, an option is to let them study worked-out examples of these learning tasks. In an MPM learning task, students have to formulate an argument – integrating their answers to some of the open-ended questions – that proves a statement to be true or false. A worked-out example of such a learning task would then demonstrate a correct argument that correctly relates and integrates the underlying propositions. In line with studies on the expertise reversal effect, performing MPM learning tasks autonomously could be effective for more knowledgeable students, while studying worked-out examples of these learning tasks could be more effective for the less knowledgeable students. As mentioned previously, the two most successful components of self-explanation appear to be filling knowledge gaps (Chi, 2000) and constructing knowledge networks (Novak, 2002). In worked-out examples, the focus may be on filling knowledge gaps more than on the construction of knowledge networks, whereas in the formulation of arguments the focus is on the construction of knowledge networks rather than on filling knowledge gaps. Less knowledgeable students can fill their knowledge gaps through the study of worked-out examples, while more knowledgeable students have sufficient knowledge to construct (and enhance) knowledge networks through argumentation. Which component of self-explanation is effective for an individual student largely depends on the individual student's prior knowledge of the subject.

For instructors who do not consider fully worked-out examples of MPM learning tasks as an option in their course, there appears to be another variant of the MPM format that may have potential for less knowledgeable students. Instead of confronting students with fully worked-out examples, one could provide them with partially worked-out examples. Both fully worked-out examples and partially worked-out examples can help students fill important knowledge gaps. A partially worked-out example of the learning task would provide students with the answers to the open-ended questions (i.e., propositions) and instruct them to formulate an argument comprising these answers.

1.7. Motivation to learn

Propositional knowledge is a necessary but not sufficient condition for conceptual understanding (Marshall, 1995). Developing conceptual understanding also requires knowledge elaboration processes like self-explanation and argumentation (Alevén & Koedinger, 2002; Fischer, 2002; Knipfer et al., 2009). To avoid disorientation on the part of the students, they should be guided into these processes of self-explanation and argumentation (Broers & Imbos, 2005; Broers, Mur, & Bude, 2005). Further, to avoid cognitive overload and motivational constraints, processes like self-explanation and argumentation should be shaped in such a way that students can directly elaborate on their prior knowledge (Kalyuga, 2009). Learning statistics is a lengthy process requiring students' motivational states and development of propositional knowledge and conceptual understanding of statistics to be taken into account. Research has shown that when one learns from an intrinsic motivation – a strong motivation from an internal desire to learn or perform – learning is more in-depth (Bruinsma, 2003), drop out is less likely, results are better (Ryan & Deci, 2000), curiosity is higher (Kuhl, 2000), one feels better in class (Levesque, Zuehlke, Stanek, & Ryan, 2004), and one is more willing to cooperate and exchange information (Martens, Gulikers, & Bastiaens, 2004). These empirically supported assumptions form the core of MPM.

1.8. Group learning

As mentioned before, some studies suggest that explanation to peers leads to more effective learning than self-explanation (Kramarski & Dudai, 2009), while other studies find no differences (Moreno, 2009) or that self-explanation leads to deeper learning than explanation to peers (Hausmann et al., 2008). One of the reasons for these mixed findings may be that the studies reported did not take into account the possibility of the expertise reversal effect. To help students who have insufficient prior knowledge, additional information is needed. This information can come from the learning task itself, from a peer student, from an instructor, or from a combination of these sources. Additional information can easily be presented in fully or partially worked-out examples. Besides, two students usually know more than one student (two students having no knowledge at all being an exception), explaining a learning task in pairs of students is an example of receiving and using additional information from a peer student. Assuming that additional information adds up to the individual student's prior knowledge (Johnson, Johnson, & Smith, 2007), it is possible that working in pairs of students is another way to compensate for insufficient prior knowledge and to reduce the expertise reversal effect. Of course, settings in which pairs of (novice) students study (partially or fully) worked-out examples together are possible.

Although MPM was originally developed as an instructional format for the individual student (Broers, 2002), suggestions have been made to apply this format in group learning settings as well. Another instructional format that has been suggested for both individual and group learning settings in complex knowledge domains is concept mapping (Novak, 2002). Compared to MPM, concept mapping may leave more freedom for discussions and may therefore be less constraining (Broers, 2009). However, an MPM learning task is supposed to force students to explicate links between propositions (Broers, 2002, 2008). If the argument falters, one can easily see whether important links are missing and/or whether many students may not be stimulated to do so by concept mapping. Further, Novak (2002, p. 552) states: "So-called concept maps that do not specify the links between 'nodes' fail to construct propositions which we see as the essential elements in representing meanings. The lack of hierarchy fails to indicate what concepts are most inclusive or most salient for a given context to which the knowledge structure is applied." Statistics is a knowledge domain comprising concepts that build on other concepts. More elementary (lower in hierarchy) concepts and propositions have to be self-explained by the students for developing an understanding of abstract ideas, and solve complex problems (higher in hierarchy) that build on these more elementary concepts and propositions. It is unclear how we express the hierarchical nature of statistical concepts and propositions in concept maps.

A common way to introduce students into the subject matter is lecturing. In fact, lectures can serve as a useful tool for increasing students' prior knowledge of the subject and to reduce the expertise reversal effect. Lecturers can present worked-out examples of learning tasks that can help students fill important knowledge gaps. Lecturers can also choose for an interactive structure: instead of presenting the subject matter in a monologue, a discussion between lecturer and students can focus on interesting problems or (partially) worked-out examples of such problems.

Another well-known instructional format for interactive learning settings is problem-based learning (PBL; Barrows, 1994, 1996; Dolmans, De Grave, Wolfhagen, & Van der Vleuten, 2005; Norman & Schmidt, 2000). Although this format may be promising once students have some

prior knowledge of a knowledge domain, the question is to what extent it can be effective for beginning students in the statistics knowledge domain. PBL is supposed to help students activate their prior knowledge. The question is, however, what is there to be activated at a stage when students still have little to no prior knowledge, or even worse: they tend to have false prior 'knowledge' (Kaplan, 2006). Furthermore, it can be questioned whether beginning students are able to formulate specific learning goals for the fundamental statistical concepts and ideas. Since the statistics knowledge domain consists of many hierarchically structured concepts, there is a need for many specific learning goals – each referring to different concepts and ideas (or procedures) – rather than a limited number of broad learning goals. Instead of learning from broad learning goals, the focus of MPM is to divide the subject matter into a limited number of short open-ended questions (i.e., to decompose the subject matter in such a way that each learning goal covers one important concept, idea or procedure) and to create learning tasks on more complex hypotheses consisting of a number of such questions.

1.9. Research questions

Based on the theoretical framework described in this chapter, six empirical studies focusing on one main research question were conducted: can MPM help students build conceptual understanding of statistics? The six empirical studies are presented in Chapters 2-7. In each study, MPM was compared and contrasted with one or more alternative instructional formats. The first three studies all focused on the student as individual learner, the last three studies focused on interactive learning rather than individual learning. Further, in the first four studies, response variables were propositional knowledge, conceptual understanding, and cognitive load. Based on the findings in these four studies and the setup of the last two studies, it was decided to focus on motivational variables instead of cognitive effort in the last two studies and to focus only on conceptual understanding as learning outcome. The research questions can be summarized as follows:

- (1) **Chapter 2:** what are the main factors affecting a student's ability to learn from performing MPM learning tasks?
- (2) **Chapter 3:** does MPM have beneficial effects on students' propositional knowledge, conceptual understanding, and cognitive load imposed on students, over and above unguided self-explanation?
- (3) **Chapter 4:** is there an interaction between the levels of students' prior knowledge and the effects of different instructional formats on students' propositional knowledge, conceptual understanding, and cognitive load imposed on students in the statistics knowledge domain?
- (4) **Chapter 5:** if prior knowledge moderates instructional effects on students' conceptual understanding, can this influence of prior knowledge be reduced by having students work in pairs?
- (5) **Chapter 6:** does MPM have beneficial effects on students' conceptual understanding of statistics and motivation to learn when applied as a lecturing method?
- (6) **Chapter 7:** can MPM be of help to increase instructional guidance and structure in PBL group discussions, and to what extent does this improve students' conceptual understanding of statistics and motivation to learn?

In the first study (**Chapter 2**), task- and student-related factors in MPM were explored. The main goal of this study was to determine a number of important factors affecting MPM success, and to provide guidelines for the design of the five experimental studies that followed. More than these experimental studies, the first study was mixed method, that is: quantitative methods and qualitative methods were combined. It was important to have an overall idea about what factors influence a student's ability to learn from such tasks and how these factors interact. To acquire as much knowledge as possible about how an instructional method works in practice, it is important to combine different research methods. The first study used a technique from the cognitive research tradition, namely students thinking aloud while performing a series of MPM learning tasks.

In the second study (**Chapter 3**), it was examined whether MPM – an instructional format in which self-explanation activity is guided – has beneficial effects in terms of propositional knowledge, conceptual understanding, and cognitive load imposed on the students when compared to unguided self-explanation. The third study (**Chapter 4**) investigated the expertise reversal effect in the domain of statistics. Low and high prior knowledge students either performed MPM learning tasks – or part of that (answering the open-ended questions) – themselves or studied worked-out examples of these learning tasks. A related question in this study was to what extent potentially beneficial effects of MPM are the result of argumentation rather than merely answering open-ended questions. In the fourth study (**Chapter 5**), the MPM condition was contrasted with a partially worked-out examples condition (instead of fully worked-out examples as in the third study), and students worked either individually or in pairs of students. Although an expertise reversal effect in terms of students' conceptual understanding was expected, a question of the fourth study was whether this effect can be reduced by having students work in pairs.

The fifth study (**Chapter 6**) focused on MPM as a potential lecturing method. Contrasting the MPM condition with a lecturing condition of less instructional guidance, the main question of this study was whether MPM has beneficial effects on students' conceptual understanding of statistics and motivation to learn when applied as a lecturing method. The sixth and final study (**Chapter 7**) contrasted MPM and PBL and explored the possibility to develop a more instructionally guided and structured form of PBL for the statistics knowledge domain. Like in the fifth study, the focus was on students' conceptual understanding of statistics and motivation to learn.

The findings of the six chapters as well as various implications for both teaching practice and further research are discussed (**Chapter 8**) and summarized (in English and in Dutch) at the end of this thesis. Finally, as chapters 2 and 4-5 have been published and chapters 3 and 6-7 have been submitted for publication, chapters 2-7 have partly overlapping theoretical framework. References and appendix materials are therefore presented at the end this thesis instead of for each chapter separately.

Chapter 2

Task-and student-related factors in propositional manipulation

Published as

Leppink, J., Broers, N. J., Imbos, Tj., Van der Vleuten, C. P. M., & Berger, M. P. F. (2011). Exploring task- and student-related factors in the method of propositional manipulation (MPM), *Journal of Statistics Education*, 19(1), [open online access]

2.1. Introduction

The statistics knowledge domain comprises abstract concepts that frequently build on other concepts and have no meaning outside the domain. This, together with other factors (e.g., the place in the study curriculum, the student's background and motivation, inappropriate instructional formats), makes it difficult for students to develop conceptual understanding of statistics (i.e., an understanding of the statistical concepts and the relationships between these concepts; Huberty et al., 1993). The method of propositional manipulation, in short MPM (Broers, 2008), aims to help students develop conceptual understanding by guiding them into self-explaining the subject matter.

2.1.1. Theoretical groundwork for MPM

When studying statistics literature, attending a lecture, or when performing a learning task on statistics, students are confronted with important concepts and core ideas. Students first have to isolate the important ideas by deriving their constituent elements, and then relate and integrate these elements into schemata and gradually develop an integrated knowledge network (Novak, 2002). Knowledge of isolated statistical ideas and elements is called propositional knowledge, whereas the ability to relate and integrate these elements is called conceptual understanding (Huberty et al., 1993). Propositional knowledge is a necessary but not sufficient condition for conceptual understanding (Marshall, 1995). Developing conceptual understanding also involves self-explanation and argumentation (Alevin & Koedinger, 2002; Fischer, 2002; Knipfer et al., 2009). In the domain of statistics, guiding students into self-explanation as in MPM appears to enhance learning outcomes more than unguided self-explanation (Broers & Imbos, 2005; Broers et al., 2005), most likely because unguided self-explanation requires students to find out themselves which are the relevant propositions in the subject matter. The latter can easily lead to disorientation on the part of the students.

Learning imposes cognitive load on students (Van Merriënboer & Sweller, 2005). Cognitive load consists of three types of load that are assumed to be additive: intrinsic load, germane load, and extraneous load. Intrinsic load depends on task complexity and the students' statistics proficiency level. This type of load should be manipulated in instructional design by selecting learning tasks that fit to the students' statistics proficiency level (Schnotz & Kürschner, 2007). As the intrinsic load imposed on students when studying statistics is usually high, a learning task that is too difficult will easily lead to cognitive overload (Kalyuga, 2009). Furthermore, all instructional features not directly beneficial for learning impose extraneous load on the student. To have sufficient capacity available for germane load, that is load from instructional features and

learning processes enhancing learning (e.g., self-explanation and argumentation), extraneous load should be minimized. Germane load is not only constrained by intrinsic and extraneous load, but also by students' interests and learning orientations, and affective and motivational aspects.

Having students learn by themselves in the domain of statistics easily leads to cognitive overload and disorientation on the part of the students, and as a consequence they will not develop proper knowledge and understanding of the subject matter. There is a need for an instructional format that stimulates the student to self-explain without experiencing cognitive overload, and this is exactly the focus of MPM.

2.1.2. MPM and domain-specific thought-processes in statistics

MPM comprises three steps. In the first step, the instructor determines the subject matter and divides it into a limited number of propositions. Propositions are statements referring to single statistical ideas and concepts (e.g., arithmetic mean, mode, and z-score). The number of propositions depends on size and content of the subject matter. As mentioned previously, intrinsic cognitive load needs to be manipulated in the instructional design by selecting learning tasks matching the students' statistics proficiency level. Therefore, which propositions are chosen by the instructor should depend on the students' statistics proficiency level. The instructor formulates questions, each referring to one proposition. Examples of propositions and questions referring to these propositions are presented in Box 2.1.

Box 2.1.

Example of propositions and questions referring to propositions

Proposition 1: *a z-score is the deviation of a score from the arithmetic mean, relative to the standard deviation.*

Question referring to proposition 1: what is a z-score?

Proposition 2: *the z-score of a score equal to the arithmetic mean equals zero.*

Question referring to proposition 2: what is the case when z equals zero?

Proposition 3: *the arithmetic mean is strongly influenced by scores in the tail of a skewed distribution.*

Question referring to proposition 3: why is the arithmetic mean not robust against skewness?

Proposition 4: *the mode in a unimodal distribution is the peak of that distribution.*

Question referring to proposition 4: what is the mode in a unimodal distribution?

Thus, the idea is that if the instructor wants students to learn the four propositions presented in Box 2.1., (s)he has to formulate questions referring to each of these propositions. By having the instructor determine and decompose the subject matter this way, students have more cognitive resources available for learning. Having the student search for the relevant propositions themselves would increase extraneous load, as this search process is not directly beneficial for learning.

In the second step of MPM, students are instructed to answer the questions formulated in the first step. Students are provided with the questions (e.g., ‘what is expressed by a z-score?’), not the actual propositions (e.g., “a z-score expresses the deviation of a score from the arithmetic mean, relative to the standard deviation”). The propositions are taught to the students in lectures and they can be found in the literature to be studied. By having students answer questions referring to propositions, they become aware of important misconceptions and they develop the propositional knowledge needed for building conceptual understanding. Students are stimulated to self-explain the subject matter and they are guided into this process of self-explanation by means of the questions. Given the abstract and cumulative nature of statistics and the frequent and tough misconceptions students have about the subject matter, this second step of MPM is a necessary step for developing conceptual understanding of statistics.

It is only in the third step of MPM that students begin to develop conceptual understanding, namely by performing a series of MPM learning tasks. In an MPM learning task, students have to relate and integrate a number of propositions into an argument that proves a given hypothesis to be either true or false. In contrast to propositions, the hypothesis typically comprises multiple statistical ideas and concepts. Therefore, hypotheses are generally of a higher complexity level than propositions. Once students master the propositions (i.e., propositional knowledge), they should relate and integrate these propositions into an argument in such a way that the argument enables them to understand why the hypothesis is true or false (i.e., conceptual understanding).

The propositions have been chosen by the instructor in the first step. For each proposition, the instructor formulates one question (for examples see Box 2.1.). In the second step, students answer these questions (and thereby discover the propositions). In the third step, the instructor gives the students a hypothesis; then the instructor attaches several propositions formed as questions to this hypothesis. The complexity level and the exact formulation of the hypothesis depend on the learning goals of the statistics course: what interrelationships between statistical ideas and concepts do we want students to know at the end of the course? Next, which questions are attached to the hypothesis depends on the learning goals of the statistics course as well as on the specific content of the hypothesis. Consider the example presented in Box 2.2.

Box 2.2.

Example of an MPM learning task in the statistics knowledge domain

Hypothesis: *If a distribution is unimodal and skewed to the right, the mode of that distribution has a negative z-score.*

- [1] Why is the arithmetic mean not robust against skewness?
- [2] What is the mode in a unimodal distribution?
- [3] What is a z-score?
- [4] What is the case when z equals zero?

Suppose, the instructor wants students to learn that although in a unimodal and symmetric distribution the mode and the arithmetic mean are equal, in a unimodal but skewed distribution the arithmetic mean is shifted towards the tail of that distribution. In a unimodal and symmetric

distribution, the z-score of the mode equals zero, whereas in a unimodal but skewed distribution the z-score of the mode is not equal to zero. Thus, the instructor formulates the hypothesis presented in Box 2.2. The hypothesis includes the concepts ‘mode’, ‘unimodal distribution’, and ‘z-score’. This explains why questions [2] and [3] have been attached. Further, the z-distribution is a unimodal and symmetric distribution, meaning that mode, median, and arithmetic mean are equal. This is why questions [1] and [4] are useful here. When students are confronted with the hypothesis only, they may not be able to answer that the hypothesis in question is true. And even if answering this question, the answer – ‘true’ or ‘false’ – may reflect a rule that was learnt by heart right before the exam, without learning the meaning of the statistical concepts the hypothesis comprises. The latter being the case, it is very likely that students will not be able to solve other hypotheses comprising the same propositions.

MPM stimulates students to engage in meaningful learning, as it stimulates them to self-explain the elements underlying the more complex hypothesis. The students must form an argument for the truth or falsity of the hypothesis based on the answers to the questions and the connections between them. Students are not expected to learn the propositions through MPM learning tasks. It is in the previous (i.e., second) step of MPM that the propositions are presented to the students, in lectures as well as in the (course) literature. Besides, questions should be formulated in such a way that they require only short answers and that each question can be related to at least one other question. Given q number of questions, the argument can comprise a maximum of $q(q - 1)/2$ pairwise connections. The number of valid connections depends on the exact content of the propositions formed as questions. In the example presented in Box 2.2., students have to create an argument comprising the answers to the four questions, meaning at least three connections and at most six connections. In an MPM learning task, students do not receive instruction on which connections should be made and which connections should be left out. The only instruction students receive is to create their argument in such a way that each question is related to at least one other question, and there is no further instruction around the learning task. A correct argument for the example is displayed in Box 2.3.

Box 2.3.

Example of an MPM argument in the statistics knowledge domain

In a unimodal distribution, the mode is the peak of the distribution [question 2]. In the case of a skewed distribution, the arithmetic mean is strongly influenced by scores in the tail of that distribution [question 1]. Therefore, the arithmetic mean in this distribution will be lying more towards the tail than the mode [questions 1 and 2 are related]. A z-score expresses how many standard deviations the original observation deviates from the arithmetic mean and in which direction [question 3]. In the case that z equals zero, the original observation does not deviate from the arithmetic mean [questions 3 and 4 related]. Given that the distribution here is skewed to the right, the mode is lower than the arithmetic mean, and therefore, the z-score of the mode is negative [questions 2 and 3]. Thus, the hypothesis is correct.

Each question refers to a single statistical idea or concept. Therefore, a correct argument comprises a relevant set of true propositions and can prove a given hypothesis to be either true or false. Which hypothesis and which questions one chooses for constructing an MPM learning

task depends on the learning goals of the statistics course as well as on the students' statistics proficiency level. For example, the learning task presented in Box 2.2. may increase understanding on the part of students who have just attended the relevant lecture and studied the accompanying literature, whereas for someone who has profound knowledge and understanding of descriptive statistics, this learning task may be too easy to increase understanding. Thus, when formulating the hypothesis and questions in an MPM learning task, students' statistics proficiency level has to be taken into account. Further, to put a learning task into a (real-life) context, contextual information (e.g., a problem) can precede the hypothesis.

Developing conceptual understanding requires a sound propositional knowledge as well as self-explanation and argumentation, and MPM encompasses all these elements. By having the instructor choose the propositions, students are guided into self-explanation of these propositions, which helps them to develop propositional knowledge. Next, it is the manipulation of propositions in learning tasks that guides students into self-explanation and argumentation on a higher, more complex level, which helps them to develop conceptual understanding.

2.1.3. Factors affecting MPM success

There are at least five factors that can affect students' ability to perform an MPM learning task and learn from such a task. First of all, lack of propositional knowledge may hamper students' ability to create an argument. As mentioned before, studying and self-explaining the propositions is a necessary step towards developing conceptual understanding. A question of interest to the current study is to what extent instructing students to study and self-explain the propositions (i.e., in the form of questions) helps them to develop propositional knowledge. Second, even if students have the propositional knowledge needed to create their argument, there is no guarantee that creating the argument contributes to learning, and whether this interacts with students' statistics proficiency level. Third, an interesting question is whether students use all their propositional knowledge in the argument explicitly, or whether they tend to leave some propositions implicit. Fourth, as in the argument every question needs to be related to at least one other proposition, choosing more propositions means students have to produce more relationships. It can be expected that cognitive load increases as the number of propositions to be integrated into the argument increases. An interesting question is the consequences of choosing more propositions in terms of learning outcomes. Fifth, depending on students' statistics proficiency level and on the complexity of the subject matter, increased cognitive load can either increase or decrease learning outcomes. Proficient students may learn optimally from an MPM learning task on relatively complex subject matter, for example from inferential statistics, or from a learning task comprising a higher number of propositions, whereas less proficient students may only benefit from an MPM learning task on less complex subject matter, for example from descriptive statistics, or from a learning task consisting of only a few propositions. The current study addressed the abovementioned factors that might affect MPM success with five research questions:

- (1) To what extent do students understand the propositions in an MPM learning task after studying the accompanying questions that refer to these propositions?
- (2) What is the effect of creating an MPM argument on cognitive load and learning outcomes?

- (3) To what extent do students integrate all propositions, represented by the questions, into their argument?
- (4) How does the number of propositions in an MPM learning task affect cognitive load and learning outcomes?
- (5) Is MPM equally effective for relatively complex subject matter (e.g., inferential statistics) as for less complex subject matter (e.g., descriptive statistics)?

Before examining MPM learning tasks in an experimental setup, it is important to have an overall idea about what factors influence a student's ability to learn from such tasks and how these factors interact. To acquire as much knowledge as possible about how an instructional method works in practice, it is important to combine different research methods. Explorative studies may precede subsequent experimental studies. Therefore, the current study was explorative, combining quantitative measures for cognitive load and qualitative measures (i.e., a mixed method approach), using a technique from the cognitive research tradition, namely thinking aloud while performing a series of MPM learning tasks.

2.2. Method

The research questions were studied by having students with different statistics proficiency levels think aloud while working on a total of six learning tasks.

2.2.1. Participants

Twenty bachelor psychology students who passed the first-year statistics exam volunteered. The first-year statistics course covered probability calculus, sampling distributions, null hypothesis testing, confidence intervals, *t*-test, one-way analysis of variance (ANOVA), and χ^2 -test. Students with an exam score of six or higher on a ten point scale pass the exam. Typically, about 55% of the students pass this exam the first time. To create sufficient variability in statistics proficiency level, two samples were drawn: ten students with a low proficiency level (i.e., low proficiency group), and ten students with a high proficiency level (i.e., high proficiency group). The lower proficiency group consisted of students who passed the exam with six or seven on a ten point scale, and the higher proficiency group consisted of students with grades eight or higher.

2.2.2. Materials

To answer the questions on the number of propositions and the potential of MPM for descriptive versus inferential statistics, the six learning tasks differed in topic and in the number of propositions to be linked in the argument. Three of the learning tasks covered different topics within descriptive statistics, including association in a two-way table (three propositions), histogram and intervals of scores (four propositions), and the usability of the correlation coefficient judging from a scatterplot (five propositions). The other three learning tasks covered different topics within inferential statistics, namely the expected mean (three propositions), Type II error (four propositions), and statistical significance (five propositions).

To answer the question to what extent students understand the propositions in an MPM learning task after studying the accompanying questions, they received a list with the questions that appeared in these learning tasks, with the instruction to study them carefully from Moore, McCabe, and Craig's (2009) textbook.

As we were interested in effects of different aspects of MPM on learning outcomes as well as on cognitive load, we needed a device to measure cognitive load imposed on the students when performing the learning tasks. Cognitive effort is an accepted indicator for cognitive load imposed on a student (Paas, 1992). Typically, students have to indicate on a nine point scale how much cognitive effort performing a task or solving a problem required from them, one being virtually no cognitive effort at all and nine being the maximum cognitive effort. Although most studies use the nine point scale, to create more variability in cognitive effort we used a visual analog scale (VAS). In the latter, students are instructed to indicate how much cognitive effort performing a task or solving a problem required by drawing a small vertical line on a horizontal continuous line going from 0 (left) to 100 (right). Such scales have been used in numerous studies in various domains the last nineteen years, and have acceptable reliability and validity for those domains (e.g., DeLoach, Higgins, Caplan, & Stiff, 1998; Myles, Troedel, Boquest, & Reeves, 1999). To our knowledge, reliability and validity estimates for studies in educational settings are still lacking. The reason why we chose the VAS in the current study was to create more variability in cognitive effort ratings. For each learning task, the VAS was administered twice. In the next section (procedure), it is explained why and how this was done.

2.2.3. Procedure

For all the learning tasks, the procedure was the following. To make sure that students understood what they had to do in a learning task, they had to first read and summarize aloud contextual information and a hypothesis. Eventual misreading or misinterpretation of language could be corrected by the researcher. At this point, students did not yet receive the questions (referring to the propositions) to be linked into an argument. Students were instructed to explain whether the hypothesis was true or false, using the contextual information at hand. Once they had given their explanation, they indicated on the VAS how much cognitive effort they experienced while performing this task. At this point, we had one solution and one cognitive effort indication (for every student) for the learning task in question. In each learning task, this solution and cognitive effort indication served as baseline measurement.

After this baseline measurement, students were confronted with the underlying questions in the learning task, each referring to one proposition. Before instructing students to create an MPM argument, students were asked to answer each of the questions presented to them (i.e., three, four, or five questions, depending on the learning task). Their answers provided information on the question to what extent students understand the propositions in an MPM learning task after studying the accompanying questions that refer to these propositions. As expected, many students still demonstrated incomplete knowledge of some of the propositions. Since we wanted to know whether students having sufficient propositional knowledge can create an MPM argument – propositional knowledge is a necessary condition for conceptual understanding – we provided all students (i.e., including those who demonstrated complete knowledge) with standard answers to the questions referring to the propositions and instructed them to create an argument integrating these propositions (i.e., answers). Their argument should prove whether the hypothesis in the learning task was true or false. Once they had created their argument, they indicated again on the VAS how much cognitive effort was experienced while performing this task. In each learning task, this solution and cognitive effort indication served as the after-treatment measurement.

Thus, in each learning task we had two solutions for every student as well as two cognitive effort indications, one after their explanation without explicitly referring to the propositions and the other after they had created their MPM argument. This enabled us to estimate the effect of creating an MPM argument on learning outcomes and cognitive load. If MPM was successful in these learning tasks, students' MPM arguments would comprise more information than their explanations given some minutes earlier. With regard to cognitive load, we did not have specific expectations, partly because of the diversity in learning tasks.

Apart from the comparison of students' MPM arguments with their explanations given before the confrontation with the questions, the MPM arguments provided information with regard to the question to what extent students explicate their knowledge of the propositions (i.e., answers to the questions) and relationships between them in their argument. To acquire additional information on the latter, every pair of questions was isolated and students had to explain what relationship they thought existed between the two questions.

2.2.4. Data analysis

The current study combined quantitative measures (i.e., with regard to cognitive load) and qualitative measures (i.e., think-aloud protocols transcribed verbatim). For the qualitative analyses, two researchers rated independently from each other. Differences in interpretation were discussed in order to seek consensus.

2.2.4.1. The effect of studying propositions

For the first research question, to what extent students understand the propositions in an MPM learning task after studying the accompanying questions that refer to these propositions, students' answers to the questions in the learning tasks were compared with the answers derived from Moore et al. (2009) and coded either correct or incorrect by two independent coders. Initial agreement between the coders was high (Cohen's $\kappa = .90$).

2.2.4.2. The effect of creating an MPM argument

For the second research question, on the effect of creating an MPM argument on cognitive load and learning outcomes, students' MPM arguments were compared with their explanations given before the confrontation with the questions. Two independent raters rated each argument as: (a) a correct and complete argument leading to a correct hypothesis evaluation (i.e., a correct 'true' or 'false'), (b) an incomplete (i.e., not all propositions integrated explicitly) and/or partly incorrect argument leading to a correct hypothesis evaluation, or (c) an incomplete and/or (partly) incorrect argument leading to an incorrect hypothesis evaluation (i.e., an incorrect 'true' or 'false'). Initial agreement between the coders was also high (Cohen's $\kappa = .81$).

Further, as in every learning task students indicated twice how much cognitive effort the learning task required from them (i.e., once before and once after creating their MPM argument), split-plot analysis of variance (ANOVA) was performed for comparing the proficiency groups with regard to the effect of creating an MPM argument on cognitive load. The level of significance α was Bonferroni corrected for each of the three p-values (i.e., one with regard to the interaction effect, and two with regard to the main effects) to correct for multiple testing and increased overall Type I error probability.

2.2.4.3. The extent to which students explicitly integrate propositions in their argument

The coding of the MPM arguments also provided information on the question to what extent students explicitly integrate propositions in their argument. However, the question that arises here is whether or not mentioning a particular relationship between two propositions reflects a mere tendency to leave out a relationship that is known by the student, or it reflects a lack of knowledge of that particular relationship on the part of the student. Therefore, it was determined per pair of propositions whether students indicated a correct relationship between the two propositions. Given that students had the answers to the questions referring to the propositions it was relatively easy for them to provide comments on the relationships.

2.2.4.4. The number of propositions in an MPM learning task

The effect of the number of propositions in a learning task on learning outcomes was examined by using the analyses of students' arguments described in section 2.2.4.2. For the effect on cognitive load, two-way within-subjects ANOVA was performed. Factors were the number of propositions and measurement point (i.e., baseline measurement being after students' explanation without the propositions to be integrated, after-treatment measurement being after students' MPM argument). The level of significance α was Bonferroni corrected for each of the three p -values.

2.2.4.5. Descriptive statistics and inferential statistics

Similar to the question on the effect of the number of propositions in a learning task, the question on the potential of MPM for descriptive versus inferential statistics was examined by using the analyses of students' arguments described in section 2.2.4.2. For the effect on cognitive load, two-way within-subjects ANOVAs were performed. Factors were subject matter (i.e., descriptive versus inferential statistics) and measurement point (i.e., baseline measurement being after students' explanation without the propositions to be integrated, after-treatment measurement being after students' MPM argument). The level of significance α was Bonferroni corrected for each of the three p -values.

2.3. Results

Each of the following paragraphs addresses the results with regard to one of the research questions of the current study.

2.3.1. The effect of studying propositions

The data suggest that students find it difficult to describe abstract statistical concepts. Although all students could answer most questions correctly and demonstrated partial knowledge with regard to those questions they could not answer correctly, a total of three questions had a very low number of correct answers in both groups. First of all, a total of fifteen students did not mention that the correlation coefficient is about linear relationship and not about just any relationship. Second, students' descriptions of the p -value revealed that many students find it difficult to interpret this value as a conditional probability. Third, although students knew examples of test statistics, they could not give a general definition of this term. Finally, eight of the ten students in the lower proficiency group confused the sampling distribution with the distribution of sample scores. In the higher proficiency group, five students made this mistake.

2.3.2. The effect of creating an MPM argument

With regard to the effect of creating an MPM argument, the data can be summarized as follows. First, when asking students to evaluate hypotheses they consider easy – because the hypothesis is self-evident, a rule learned by heart, or too easy for their statistics proficiency level – they hardly motivate their evaluations. Second, creating an MPM argument does not guarantee that students replace their misconceptions by correct knowledge. A quote from a student in the lower proficiency group:

“...eh, yes the value is close to 0 and thus I would say that there is almost no association, linear association, and, yes, I cannot see any non-linear association here, because whether you draw a straight line or any other line you will not manage because the points are so spread out and thus I suppose that the correlation coefficient, eh, yes, does give the correct value and given that this is here .03, it is, eh, indeed not a good summary of the association between x and y ”

Third, given the difficulties students experience when describing the statistical concepts mentioned, creating an MPM argument integrating these concepts is likely to be difficult for them as well. An example is the following:

“Okay, so the hypothesis was eh, whether the expected mean can be expressed in a number... hmm... now I am confused... eh... okay... so, I have the, eh, mean of the, eh, neuroticism scores... hmm, so the distribution of the sample is not the same as the sampling distribution... hmm, yes, but the distribution of the sample... sample mean is eh... I just said that the hypothesis is true and I still think that eh... well the sample mean is the expected mean, or not? The, sorry, eh, if the sample is drawn at random, then they must be equal... [researcher reminds the student that all propositions have to be integrated into the argument, conform the instructions] ... hmm, yes, I, I see no relation here between the questions... I have only the distribution of the sample and not the sampling distribution... so eh... that cannot be?... eh but then the hypothesis is incorrect... [researcher reminds the student that all propositions have to be integrated into the argument, conform the instructions] ... hmm, yes, here I have only the mean of the sample, so, eh, that is just a one value of... eh, it is just one sample... [researcher reminds the student that all propositions have to be integrated into the argument, conform the instructions] eh, the sample mean is not equal to the expected mean, and I do not have the population mean here... yes... the population mean equals the expected mean... so the hypothesis is false, because the population mean is not given and that one is equal to the expected mean, now I cannot compute the population mean, because I just have the distribution of one sample, and thus not the sampling distribution of the mean when you repeat an infinite number of times, because, eh, de expected mean is the mean of the sampling distribution and that one is not given here... [researcher: thus the hypothesis is?] ... false.”

This example is from a student in the higher proficiency group. Students' arguments in this learning task illustrate that frequent misconceptions about the expected mean are that the expected mean equals the sample mean or that the expected mean can be computed from mere sample data. Before the instruction to create an MPM argument in this learning task, sixteen of the twenty students – eight in each group – gave an incorrect explanation leading to an incorrect hypothesis evaluation. The instruction to create an MPM argument made six of the proficient students aware of their mistake, and as a result they came to a correct hypothesis evaluation. In the lower proficiency group, this change was limited to one student. This finding is in line with

the finding reported in the previous section that describing the concept of sampling distribution is difficult, especially for the less proficient students.

With regard to cognitive load, on average, proficient students reported lower cognitive load than their less proficient peers. Table 2.1. displays the average cognitive efforts and standard deviations (*SD*) for both proficiency groups before and after the instruction to create an MPM argument, for the three learning tasks on descriptive statistics.

Table 2.1.
Means (and *SD*) of cognitive effort required for the learning tasks on descriptive statistics

Condition	<i>N</i>	<i>M</i> (<i>SD</i>)
Explanation before the questions		
Lower proficiency group	10	45.01 (17.73)
Higher proficiency group	10	31.79 (17.39)
MPM argument		
Lower proficiency group	10	39.40 (13.41)
Higher proficiency group	10	31.52 (16.58)

The group by condition interaction was not significant, [$F(1, 18) = 0.336, p > .50, \eta^2 = .018$] and the same was the case for the main effect of condition, [$F(1, 18) = 0.408, p > .50, \eta^2 = .022$], as well as for the group effect, [$F(1, 18) = 3.43, p = .08, \eta^2 = .160$]. As the effect sizes indicate a large size effect for the group effect, absence of statistical significance is probably due to small sample sizes. Table 2.2. displays the average cognitive efforts and standard deviations for both proficiency groups before and after the instruction to create an MPM argument, for the three learning tasks on inferential statistics.

Table 2.2.
Means (and *SD*) of cognitive effort required for the learning tasks on inferential statistics

Condition	<i>N</i>	<i>M</i> (<i>SD</i>)
Explanation before the questions		
Lower proficiency group	10	45.96 (12.00)
Higher proficiency group	10	36.34 (17.35)
MPM argument		
Lower proficiency group	10	54.41 (13.41)
Higher proficiency group	10	46.23 (16.40)

The group by condition interaction was not significant, [$F(1, 18) = 0.048, p > .80, \eta^2 = .003$], the main effect of condition was significant, [$F(1, 18) = 7.729, p < .05, \eta^2 = .300$], and the group effect was not significant, [$F(1, 18) = 2.340, p > .10, \eta^2 = .115$]. As the effect sizes indicate a medium to large size effect for the group effect, absence of statistical significance is probably due

to small sample sizes. The data suggest that evaluating a hypothesis by means of an MPM argument imposes additional cognitive load on students when learning inferential statistics, but not when learning descriptive statistics.

2.3.3. *The extent to which students explicitly integrate propositions in their argument*

Despite the instruction to explicitly integrate propositions, students tend to restrict themselves to merely listing the propositions. In most cases, it was only after repeated asking by the researcher to relate and integrate propositions that students attempted to do so. The researchers anticipated this possibility, and therefore after creating the MPM argument, students were instructed per pair of underlying propositions what the relationship between the two propositions is. Although initially not all students were able to give appropriate answers to all questions, they were now able to describe the relationships for nearly all pairs of propositions. Not clear is whether this ability was a consequence of seeing the standard answers to the questions or partly an effect of creating an MPM argument. These findings and those in the previous two sections together suggest two practical implications for MPM. First, sufficient propositional knowledge is a firm necessary condition for creating an MPM argument. Second, to avoid that relevant propositional knowledge does not appear in the argument, the instruction to integrate all propositions in the learning task in the argument should be clear and repeated.

2.3.4. *The number of propositions in an MPM learning task*

Table 2.3. presents the average cognitive efforts and standard deviations for the number of propositions (i.e., for both contents and for all twenty students) before and after the instruction to create an MPM argument.

Table 2.3.
Means (and *SD*) of cognitive effort experienced as a function of the number of propositions

Condition	<i>N</i>	<i>M (SD)</i>
Explanation before the questions		
Three propositions	20	49.03 (19.21)
Four propositions	20	37.90 (16.75)
Five propositions	20	32.40 (14.57)
MPM argument		
Three propositions	20	55.41 (13.27)
Four propositions	20	34.76 (21.06)
Five propositions	20	38.51 (17.82)

Two-way within-subjects ANOVA revealed a non-significant interaction between the number of propositions and the learning format, [$F(2, 38) = 2.405, p > .10, \eta^2 = .112$], and a significant main effect for the number of propositions, [$F(2, 38) = 15.552, p < .001, \eta^2 = .450$]. From further analysis it appears that the proposition effect is quadratic, [$F(1, 18) = 6.229, p < .05, \eta^2 = .247$] and that a learning task consisting of three underlying propositions requires significantly more cognitive effort than a learning task consisting of four or five underlying propositions. It is difficult

to interpret this quadratic effect, as it may indicate some kind of novelty effect: for both descriptive and inferential statistics, the first learning task students were confronted with consisted of three propositions and the learning tasks following consisted of four and five propositions respectively.

2.3.5. Descriptive statistics and inferential statistics

Table 2.4. presents the average cognitive efforts and standard deviations for statistical topic before and after the instruction to create an MPM argument.

Table 2.4.
Means (and *SD*) of cognitive effort required as a function of statistical topic

Condition	<i>N</i>	<i>M (SD)</i>
Explanation before the questions		
Descriptive statistics	20	38.40 (18.39)
Inferential statistics	20	41.15 (15.23)
MPM argument		
Descriptive statistics	20	35.46 (15.33)
Inferential statistics	20	50.32 (15.17)

The average cognitive effort to perform the assignment was not only higher for inferential assignments than for descriptive assignments, a difference in trend for the different subject matters exists as well. Two-way within-subjects ANOVA revealed a significant interaction between content and learning format, [$F(1, 19) = 5.057, p < .05, \eta^2 = .210$]. Subsequent analysis made clear that the difference between conditions was significant for the learning tasks on basic inferential statistics, [$F(1, 19) = 8.140, p < .01, \eta^2 = .300$], and that creating an MPM argument required significantly more cognitive effort in learning tasks on elementary inferential statistics than in learning tasks on descriptive statistics, [$F(1, 19) = 9.730, p < .005, \eta^2 = .339$]. Apparently, it demands a considerable cognitive effort from the student to do an inferential statistics assignment in MPM format. Together with the finding that, in the higher proficiency group, confronting students in MPM format with the concepts of sampling distribution and expected mean leads to an increased awareness of a misconception and, as a consequence, better performance, one can infer that the additional cognitive effort due to working in MPM format can lead to better learning outcomes, at least for those having a certain proficiency in (inferential) statistics.

2.4. Discussion

The first major finding is that instructing students to study a list of propositions is effective for the easier propositions but less effective for propositions referring to abstract statistical concepts. The instruction to study the propositions should be such that students spend more time studying abstract statistical concepts. This could be achieved by a list of propositions in which abstract concepts are represented with more propositions than easier concepts. In the

current study, each concept was represented by just one proposition. Future studies may examine the effect of more than one proposition referring to an abstract concept. Sufficient propositional knowledge is a necessary condition for creating an MPM argument. Therefore, if representation by abstractness in the propositions list leads to more propositional knowledge, the quality of MPM arguments may be increased as well.

With regard to the effect of creating an MPM argument, the current study reveals two extremes. On the one hand, creating an MPM argument integrating abstract statistical concepts is difficult and does not necessarily lead to misconception awareness with regard to these concepts. On the other hand, when asking students to evaluate hypotheses they consider easy – because the hypothesis is self-evident, a rule learned by heart, or too easy for their statistics proficiency level – they hardly motivate their evaluations, nor do they feel motivated to create an MPM argument. In line with findings in previous studies (Kalyuga et al., 2003; Wetzels, 2009), the complexity level of each of the aspects of an MPM learning task – contextual information, hypothesis and underlying propositions – must be in accordance with the students' statistics proficiency level. A very easy learning task is not likely to stimulate students to create an MPM argument, whereas a very difficult learning task may lead to cognitive overload for the students. Future studies should focus on the development and validation of MPM learning tasks for different statistics proficiency levels.

Despite the instruction to explicitly integrate propositions, students tend to restrict themselves to merely listing the propositions. Although in some learning tasks, this behavior is most likely the consequence of a discrepancy between the students' statistics proficiency level and task difficulty, future studies could focus on the effectiveness of different forms of the instruction to integrate all propositions in the argument. For example, to avoid omission of relevant propositional knowledge in the argument, one could repeat the instruction every time a link between two propositions is made by the student. Another type of instruction is to formulate more MPM arguments in a particular learning task.

Since we did not counterbalance the order of learning tasks for the students, the findings with regard to the number of propositions in an MPM learning task do not enable us to draw straightforward conclusions, only that it appears important to make the student familiar with MPM. The findings with regard to cognitive effort appear to reflect that once the student is familiar with MPM, the instructional format itself imposes less cognitive load on the student. To examine the effect of the number of propositions in an MPM learning task, future studies should treat this effect as a between-subjects factor or counterbalance the order of the learning tasks.

The comparison of descriptive statistics and inferential statistics indicates that the topic itself needs to be complex enough to stimulate students to work according to MPM. However, the finding that the students in the current study are more stimulated to create MPM arguments for inferential statistics than for descriptive statistics may reflect the aforementioned discrepancy between the students' statistics proficiency level and task difficulty. In their study curriculum, a course on descriptive statistics precedes the course on inferential statistics and the students were selected based on their exam score for the latter. Therefore, future studies should examine the potential of MPM for learning descriptive statistics among students starting the course on descriptive statistics.

The current study has a few limitations. First of all, the results are based on small sample sizes and depend upon the actual MPM learning tasks created for this study. Different hypotheses

based on more fundamental concepts with more basic questions might lead the students to a better or worse understanding than the actual MPM learning tasks. Second, task-specific characteristics are a confounding factor in some of the comparisons made in the current study. Future studies might use assignments that are hierarchical, that is, different assignments on one specific topic only varying in the number of constituent propositions. The number of propositions should then be a between-subjects factor or the order of the learning tasks should be counterbalanced. A third limitation of the current study is that all students were first instructed to explain in their own words whether the hypothesis in the learning task was true or false and subsequently had to create an MPM argument, based on the standard answers to the questions referring to the propositions. It is possible that the learning effect that was found in some of the learning tasks was partly induced by the fact that students had been working some time already on the learning task and that they had full knowledge of the propositions (i.e., in the form of the standard answers). A fourth limitation is that students in the current study received a single prompt to create an MPM argument. The results indicate that this procedure can help students whose proficiency or prior knowledge matches the difficulty level of the learning task to become aware of a misconception and develop conceptual understanding (i.e., by self-explaining and integrating the constituent propositions). Taking the third and fourth limitation of the current study together, in future studies the instructional format (i.e., explaining in their own words and creating an MPM argument) should be a between-subjects factor consisting of three conditions, namely a control condition in which students explain in their own words, an MPM condition in which the student receives a single prompt to create an MPM argument and a third condition in which the student receives a number of prompts to work this way or receives the instruction to try to formulate more MPM arguments for the same learning task.

The students in the current study did not have a stake in the outcome. A possible consequence is that they may not have tried as hard as students would have if they had a stake in the outcome (e.g., a grade on their results). Therefore, a fifth limitation may be students' learning outcomes in our study were lower than learning outcomes in the educational practice. Finally, as the students in the current study were selected on the basis of their performance on the subject matter, they were not novices. On the one hand, given the findings with regard to the students' proficiency in statistics, some studies should contrast the instructional formats for a larger number of tasks, varying in difficulty level, for different levels of proficiency or prior knowledge. On the other hand, since MPM aims to both guide novices into a complex knowledge domain like statistics and help other students to develop a better (conceptual) understanding of the domain, other studies should use novices (i.e., no prior knowledge about the topic whatsoever) as participants in order to determine whether they can profit from MPM as well. In the latter case, the learning tasks to be constructed need to be relatively easy, since the current study has shown that MPM can only be fruitful if the learning tasks match the students' prior knowledge of the topic.

MPM is a concrete instructional method for the statistics knowledge domain. In this article, we present task- and student-related factors influencing students' ability to learn from an MPM learning task (i.e., statistics proficiency level, subject matter, the number of propositions in the learning task, and the instructions). It is now important to examine each of these factors in subsequent experimental studies.

Chapter 3

The effect of guiding students into self-explanation

Submitted for publication

3.1. Introduction

Statistics is an important subject in many university curricula. Although students usually develop some knowledge of definitions, statistical principles and basic ideas (i.e., propositional knowledge; Broers, 2002), they often lack the ability to interrelate and structure their knowledge (i.e., conceptual understanding; Broers, 2009). The current article addresses the problem of how self-explanation, and guiding self-explanation, may help students improve their knowledge and understanding of statistics. Self-explanation is a form of constructive learning that has been proven to enhance knowledge and understanding of learning material more than merely reading the learning material (i.e., passive learning), highlighting passages of text or repeating text sentences verbatim (i.e., active learning; Fonseca & Chi, 2011). Moreover, when compared to other constructive learning strategies (e.g., generative summarizing, concept mapping, and drawing diagrams), self-explanation appears to give superior learning outcomes. Although in the latter comparison effect sizes vary from large (Roscoe & Chi, 2008) to very small (King, 1992), beneficial self-explanation effects on learning and problem-solving have been demonstrated across a variety of domains and settings (Atkinson et al., 2003). The two most successful cognitive mechanisms of self-explanations appear to be filling knowledge gaps (Chi, 2000) and constructing knowledge networks (Novak, 2002).

3.1.1. Beneficial self-explanation effects

Although beneficial self-explanation effects have been demonstrated in different domains and settings, the question is whether novice students have sufficient prior knowledge and argumentation skills to learn from explaining the subject matter (Kalyuga, 2009). A recent think-aloud study by Leppink, Broers, Imbos, Van der Vleuten, and Berger (2011a; Chapter 2 of this thesis) illustrates that students' prior knowledge about the subject matter moderates the extent to which students learn from self-explanation. Cognitive load theory provides an explanation for this finding. Learning imposes cognitive load on students (Van Merriënboer & Sweller, 2005). Working memory is limited in capacity as well as in duration (Miyake & Shah, 1999). When consciously processing new elements of information, which in complex knowledge domains like statistics are often interrelated, working memory can be overloaded (Kalyuga, 2009). Cognitive load theory assumes that the available knowledge structures in long-term memory (i.e., prior knowledge) are essential for preventing working memory overload as well as for guiding cognitive processes when learning (Van Merriënboer & Sweller, 2005). Cognitive load imposed on students should therefore be in accordance with their prior knowledge.

Cognitive load consists of three types of load that are assumed to be additive: intrinsic load, germane load, and extraneous load. Intrinsic load depends on task complexity and students' prior knowledge about the subject matter. This type of load should be manipulated in instructional design by selecting learning tasks that match students' prior knowledge (Kalyuga, 2009). Doing

so, intrinsic load is matched to students' prior knowledge. If intrinsic load is too low, this means that the learning task and/or the subject is too easy for the student, and as a consequence, students will not learn from the learning task at hand. If the intrinsic load is too high, this means that either the subject is too complex or the learning task is too difficult for the student, and hence, the student does not have sufficient capacity available to engage in germane load activities. Germane load arises from instructional features that stimulate cognitive processes beneficial for learning, whereas all instructional features not directly beneficial for learning impose extraneous load on students. As the intrinsic load imposed on students when studying statistics is usually high, extraneous load should be minimized to avoid cognitive overload (Kalyuga & Hanham, 2010). By minimizing extraneous load and matching intrinsic load to the students' prior knowledge, students can engage in self-explanation, and this imposes germane load on students.

In the ideal situation, what in cognitive load theory is called germane load activities (i.e., learning activities that are beneficial for learning), is what Bjork (1994) defines as desirable difficulties in learning. Desirable difficulties are manipulations in instructional design intended to introduce difficulties during study in such a way that long-term learning is enhanced. However, Kornell and Bjork (2007, p. 221) note: "The very fact that desirable difficulties introduce challenges and can decrease a student's perceived rate of learning may lead students to avoid rather than select such strategies." Novice students may perceive that their limited prior knowledge is insufficient to learn from self-explanations and, as a consequence, they may simply not engage in self-explanation. It appears that self-explanation activities should be well-planned and well-structured, and they should be tailored to students' prior knowledge, so that students perceive they are able to learn from self-explaining and that they do not experience cognitive overload when actually engaging in this process. By planning, structuring, and tailoring self-explanation activities to students' prior knowledge, they can engage in knowledge elaboration (Kalyuga, 2009). Knowledge elaboration is using prior knowledge to continuously expand, organize, restructure, interconnect, and integrate new elements of information. Knowledge elaboration processes are cognitive processes that impose germane load on the students.

Engaging in self-explanation is a time consuming activity, and it may stimulate students to engage in other knowledge elaboration processes, for example argumentation. Both self-explanation (Aleven & Koedinger, 2002; Atkinson, et al., 2003) and argumentation (Fischer, 2002; Knipfer et al., 2009) can enhance learning and problem-solving skills in complex knowledge domains. However, it is only if students are given sufficient time that – in terms of learning outcomes and cognitive effort required – learning by self-explanation is more effective than learning methods like observational learning, inquiry learning, and hypermedia learning (Eysink et al., 2009).

Besides, different studies have shown that self-explanation enhanced by prompting is more effective than spontaneous self-explanation (Atkinson et al., 2003; Chi et al., 1994). For example, in the study by Chi and colleagues (1994), twenty-four students were given a biology text on the human circulatory system. Of these, fourteen students received a prompt to self-explain after reading each individual line of the text. The control group, consisting of the remaining ten students, simply received the instruction to read the same text twice. Administration of a posttest demonstrated that the self-explanation group had made more progress than the

controls and especially performed better on questions that required knowledge inferences and use of common sense knowledge.

Moreover, it appears that guiding or assisting self-explanation prompts by means of open questions is more effective than merely prompting self-explanation (Berthold et al., 2009). An explanation for the latter is that when students are not guided into self-explanation, they need to search the relevant subject matter themselves before they can self-explain, and this can easily lead to disorientation and therefore less efficient learning (Eysink et al., 2009). Since students guided into self-explanation do not need to search the relevant subject matter themselves, they have more time and capacity available for self-explanation and argumentation. Finally, since research suggests that separating successive studying sessions rather than massing such sessions (i.e., spacing; Bjork, 1994; Kornell & Bjork, 2007) has positive effects on long-term learning (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1988), it appears recommendable to plan self-explanation at different stages in the learning process, for example in successive sessions in a statistics course.

3.1.2. Guiding self-explanation by manipulating propositions

When studying the statistics literature, attending a lecture, or performing a learning task, students are confronted with important concepts and core ideas. Students first have to isolate the important ideas by deriving their constituent elements, and then relate and integrate these elements into schemata and gradually develop an integrated knowledge network (Novak, 2002). Broers (2002) developed a three-step instructional method – the method of propositional manipulation (MPM) – that is supposed to help students develop a proper understanding of statistics by guiding them into self-explaining the subject matter. The aforementioned study by Leppink and colleagues (2011a; Chapter 2 of this thesis) suggests that propositional knowledge is a necessary but not sufficient condition for conceptual understanding of statistics. Developing conceptual understanding also involves self-explanation and argumentation (Alevén & Koedinger, 2002; Fischer, 2002; Knipfer, et al., 2009). Quasi-experimental studies suggest that, in the domain of statistics, guiding students into self-explanation as in MPM appears to enhance propositional knowledge (Broers & Imbos, 2005) and conceptual understanding (Broers et al., 2005) more than unguided self-explanation. A plausible explanation of this appears that unguided self-explanation requires students to find out themselves which are the relevant propositions in the subject matter. The latter can easily lead to disorientation on the part of the students and, as a consequence, they will not develop proper knowledge and understanding of the subject matter. There is a need for an instructional format that stimulates the student to self-explain without experiencing cognitive overload, and this is exactly the focus of MPM.

MPM comprises three steps. In the first step, the instructor determines the subject matter and divides it into a limited number of propositions. Propositions are statements referring to single statistical ideas and concepts (e.g., arithmetic mean, mode, and z-score). The number of propositions depends on size and content of the subject matter. As mentioned previously, intrinsic cognitive load needs to be manipulated in the instructional design by selecting learning tasks matching the students' statistics proficiency level. Therefore, which propositions are chosen by the instructor should depend on the students' statistics proficiency level. The instructor formulates questions, each referring to one proposition. Examples of propositions and questions referring to these propositions are presented in Box 3.1.

Box 3.1.

Example of propositions and questions referring to propositions

Proposition 1: *a z-score is the deviation of a score from the arithmetic mean, relative to the standard deviation.*

Question referring to proposition 1: what is a z-score?

Proposition 2: *the z-score of a score equal to the arithmetic mean equals zero.*

Question referring to proposition 2: what is the case when z equals zero?

Proposition 3: *the arithmetic mean is strongly influenced by scores in the tail of a skewed distribution.*

Question referring to proposition 3: why is the arithmetic mean not robust against skewness?

Proposition 4: *the mode in a unimodal distribution is the peak of that distribution.*

Question referring to proposition 4: what is the mode in a unimodal distribution?

Thus, the idea is that if the instructor wants students to learn the four propositions presented in Box 3.1., (s)he has to formulate questions referring to each of these propositions. By having the instructor determine and decompose the subject matter this way, students have more cognitive resources available for learning. Having the student search for the relevant propositions themselves would increase extraneous load, as this search process is not directly beneficial for learning.

In the second step of MPM, students are instructed to answer the questions formulated in the first step. Students are provided with the questions (e.g., 'what is expressed by a z-score?'), not the actual propositions (e.g., 'a z-score expresses the deviation of a score from the arithmetic mean, relative to the standard deviation'). The propositions are taught to the students in lectures and they can be found in the literature to be studied. By having students answer questions referring to propositions, they become aware of important misconceptions and they develop the propositional knowledge needed for building conceptual understanding. Students are stimulated to self-explain the subject matter and they are guided into this process of self-explanation by means of the questions. Given the abstract and cumulative nature of statistics and the frequent and tough misconceptions students have about the subject matter, this second step of MPM is a necessary step for developing conceptual understanding of statistics.

It is only in the third step of MPM that students begin to develop conceptual understanding, namely by performing a series of MPM learning tasks. In an MPM learning task, students have to relate and integrate a number of propositions into an argument that proves a given hypothesis to be either true or false. In contrast to propositions, the hypothesis typically comprises multiple statistical ideas and concepts. Therefore, hypotheses are generally of a higher complexity level than propositions. Once students master the propositions (i.e., propositional knowledge), they should relate and integrate these propositions into an argument in such a way that the argument enables them to understand why the hypothesis is true or false. The latter requires conceptual understanding of the subject matter.

The propositions have been chosen by the instructor in the first step. For each proposition, the instructor formulates one question (for examples see Box 3.1.). In the second step, students answer these questions (and thereby discover the propositions). In the third step, the instructor gives the students a hypothesis; then the instructor attaches several propositions formed as questions to this hypothesis. The complexity level and the exact formulation of the hypothesis depend on the learning goals of the statistics course: what interrelationships between statistical ideas and concepts do we want students to know at the end of the course? Next, which questions are attached to the hypothesis depends on the learning goals of the statistics course as well as on the specific content of the hypothesis. Consider the example presented in Box 3.2.

Box 3.2.

Example of an MPM learning task in the statistics knowledge domain

Hypothesis: *If a distribution is unimodal and skewed to the right, the mode of that distribution has a negative z-score.*

- [1] Why is the arithmetic mean not robust against skewness?
- [2] What is the mode in a unimodal distribution?
- [3] What is a z-score?
- [4] What is the case when z equals zero?

Suppose, the instructor wants students to learn that although in a unimodal and symmetric distribution the mode and the arithmetic mean are equal, in a unimodal but skewed distribution the arithmetic mean is shifted towards the tail of that distribution. In a unimodal and symmetric distribution, the z-score of the mode equals zero, whereas in a unimodal but skewed distribution the z-score of the mode is not equal to zero. Thus, the instructor formulates the hypothesis presented in Box 3.2. The hypothesis includes the concepts ‘mode’, ‘unimodal distributions’, and ‘z-score’. This explains why questions [2] and [3] have been attached. Further, the z-distribution is a unimodal and symmetric distribution, meaning that mode, median, and arithmetic mean are equal. This is why questions [1] and [4] are useful here. When students are confronted with the hypothesis only, they may not be able to answer that the hypothesis in question is true. And even if answering this question, the answer – ‘true’ or ‘false’ – may reflect a rule that was learnt by heart right before the exam, without learning the meaning of the statistical concepts the hypothesis comprises. The latter being the case, it is very likely that students will not be able to solve other hypotheses comprising the same propositions.

MPM stimulates students to engage in meaningful learning, as it stimulates them to self-explain the elements underlying the more complex hypothesis. The students must form an argument for the truth or falsity of the hypothesis based on the answers to the questions and the connections between them. Students are not expected to learn the propositions through MPM learning tasks. It is in the previous (i.e., second) step of MPM that the propositions are presented to the students, in lectures as well as in the (course) literature. Besides, questions should be formulated in such a way that they require only short answers and that each question can be related to at least one other question. Given q number of questions, the argument has to

comprise a minimum of $q - 1$ and can comprise a maximum of $q(q - 1)/2$ pairwise comparisons. The number of valid connections depends on the exact content of the propositions formed as questions. In the example presented in Box 3.2., students have to create an argument comprising the answers to the four questions, meaning at least three connections and at most six connections. In an MPM learning task, students do not receive instruction on which connections should be made and which connections should be left out. The only instruction students receive is to create their argument in such a way that each question is related to at least one other question, and there is no further instruction around the task. A correct argument for the example is displayed in Box 3.3.

Box 3.3.

Example of an MPM argument in the statistics knowledge domain

In a unimodal distribution, the mode is the peak of the distribution [question 2]. In the case of a skewed distribution, the arithmetic mean is strongly influenced by scores in the tail of that distribution [question 1]. Therefore, the arithmetic mean in this distribution will be lying more towards the tail than the mode [questions 1 and 2 are related]. A z-score expresses how many standard deviations the original observation deviates from the arithmetic mean and in which direction [question 3]. In the case that z equals zero, the original observation does not deviate from the arithmetic mean [questions 3 and 4 related]. Given that the distribution here is skewed to the right, the mode is lower than the arithmetic mean, and therefore, the z-score of the mode is negative [questions 2 and 3]. Thus, the hypothesis is correct.

Each question refers to a single statistical idea or concept. Therefore, a correct argument comprises a relevant set of true propositions and can prove a given hypothesis to be either true or false. Which hypothesis and which questions the instructor chooses for constructing an MPM learning task depends on the learning goals of the statistics course as well as on the students' prior knowledge of the subject matter. For example, the learning task presented in Box 2 may increase understanding on the part of students who have just attended the relevant lecture and studied the accompanying literature about this topic, whereas for someone who has profound knowledge and understanding of descriptive statistics, this learning task may be too easy to increase understanding. Thus, when formulating the hypothesis and questions in an MPM learning task, students' prior knowledge of statistics has to be taken into account. MPM learning tasks should be tailored to students' prior knowledge about the subject matter to be studied. Further, to put a learning task into a (real-life) context, contextual information (e.g., a problem) can precede the hypothesis.

Developing conceptual understanding requires a sound propositional knowledge as well as self-explanation and argumentation, and MPM encompasses all these elements. MPM is supposed to guide students into self-explanation at two different levels, a lower level and a higher level. Before performing MPM learning tasks, students have to answer the questions referring to isolated propositions. These answers can be given in the form of short propositions, statements that are considered always true (e.g., "the correlation measures the direction and strength of the linear relationship between two quantitative variables"; Moore et al., 2009, p. 102). This is self-explanation at the lower level. By having the instructor choose the propositions,

students are guided into self-explanation of these propositions, which helps them to develop propositional knowledge. Next, it is the manipulation of propositions in learning tasks that guides students into self-explanation and argumentation of a higher, more complex level. Performing the MPM learning tasks involves self-explanation at a higher level, because these learning tasks instruct the students to interrelate and structure the propositional knowledge derived at the lower level. It is in this way that MPM is supposed to help students to develop conceptual understanding of statistics.

3.1.3. The current experiment

Although quasi-experimental studies suggest that MPM has beneficial effects on propositional knowledge (Broers & Imbos, 2005) and conceptual understanding (Broers, et al., 2005), these studies failed to answer the question how to interpret this self-explanation effect, and to what extent beneficial self-explanation effects can be explained by time-on-task, that is: the time students spend on the study matter. Neither was time-on-task kept constant, nor was it included as a covariate in the analyses of these studies. Moreover, none of the previous studies focused on the question to what extent guiding students into self-explanation influences cognitive load imposed on them when learning and when applying their knowledge in an exam situation. Even if an instructional method has potentially beneficial effects, it is important to know how much effort it takes for students (i.e., how much cognitive load is imposed on them) to learn according to this method. Finally, an instructional method requiring more effort may decrease cognitive load during exam situations, if learning according to this method enables students to enhance their knowledge structures in long-term memory (Van Merriënboer & Sweller, 2005).

The current study addressed all the questions raised above, by comparing MPM with a merely prompting (i.e., not guiding) self-explanation condition and a reading-only condition (i.e., no prompt to self-explain) as a control group. Although students in the control group did not receive an explicit prompt to self-explain, in the Methods and Results section of this paper, it becomes clear that students in this condition had sufficient opportunity to study their text in a way they were used to. Students were given this opportunity to make the control condition less artificial and slightly more representative for real studying behavior. Nevertheless, we call this condition the reading condition, because students were instructed to read the text over and over again, and thus, reading was the main activity in this condition. Since in MPM self-explanation activity is guided, merely comparing the reading control condition and the MPM condition would not enable us to separate self-explanation effects and additional effects of guiding self-explanation. Therefore, we need a second control condition, in which self-explanation is not guided. Various studies suggest that unguided self-explanation enhances learning outcomes more than not engaging in self-explanation (Atkinson, et al., 2003; Chi, et al., 1994). Although the aforementioned quasi-experimental studies (Broers & Imbos, 2005; Broers, et al., 2005) suggest that guiding self-explanation enhances knowledge and understanding of statistics more than merely prompting students to self-explain (i.e., unguided self-explanation) or not prompting self-explanation at all (i.e., no self-explanation, or in some cases perhaps spontaneous self-explanation), the current study addressed the questions that these studies did not address (i.e., how to interpret self-explanation effects and what exactly is the added value of guiding self-explanation, the influence of time-on-task, and cognitive load), and used an experimental setup.

In view of the difficulties higher education students generally encounter when studying statistics, controlled research on the effect of instructional methods in this domain is highly relevant.

Because MPM was developed for the statistics knowledge domain, the current study focused on this knowledge domain. Four hypotheses were tested, each of them suggesting that potentially beneficial effects of MPM reflect both a self-explanation effect and an additional effect of guiding self-explanation rather than a mere self-explanation effect. In an MPM learning task, students are confronted with a set of relevant and interrelated questions referring to propositions, and performing the learning task can help them to develop conceptual understanding. When merely prompting (i.e., not guiding) self-explanation, students first need to search for relevant and interrelated questions themselves, and this can lead to elevated cognitive load and meanwhile less enhancement in knowledge and understanding of statistics (Eysink et al., 2009). Thus, the following four hypotheses were tested:

- (1) When studying basic inferential statistics, MPM imposes less cognitive load on students than prompting self-explanation or not prompting self-explanation at all.
- (2) When applying knowledge in an exam situation, MPM imposes less cognitive load on students than prompting self-explanation or not prompting self-explanation at all (in other words: students who learn via MPM experience lower cognitive load during the exam).
- (3) MPM enhances propositional knowledge more than prompting self-explanation or not prompting self-explanation at all.
- (4) MPM enhances conceptual understanding more than prompting self-explanation or not prompting self-explanation at all.

3.2. Method

The current experiment investigated the effect of guiding self-explanation on cognitive load, propositional knowledge, and conceptual understanding of statistics, and whether prior knowledge moderates this effect.

3.2.1. Participants and experimental design

Fifty first-year medicine students who had not yet attended any university statistics course were allocated at random to one of three possible conditions: a reading control condition ($n = 17$), a merely prompting (i.e., unguided) self-explanation condition ($n = 16$), and the MPM (i.e., guided self-explanation) condition ($n = 17$).

3.2.2. Materials

Materials were: (1) a prior knowledge test on statistical reasoning (i.e., a subset from the Statistical Reasoning Assessment; Garfield, 2003); (2) a text of four pages on basic inferential statistics, composed by the authors of the manuscript from chapters 4-6 of Moore et al. (2009), that had been subjected to a pilot-study for assessing its difficulty level and time required to read it properly; (3) one study task per group; (4) a Dutch validated version of Paas' (1992) nine-point scale for measuring cognitive load; (5) a 50 minutes test consisting of one part measuring propositional knowledge and one part measuring conceptual understanding.

3.2.3. Procedure

After the prior knowledge test on statistical reasoning, students were presented the text on basic inferential statistics and they were instructed to work for 60 minutes. Thus, time-on-task was kept constant. Another approach to controlling for time-on-task is to record students' actual time on task and to include time-on-task as a covariate in your model. Given the relatively small study assignment and given that even students in the reading condition made notes, highlighted parts of the study text, and used almost the full studying time, the latter approach might have led to severe restriction of range in time-on-task in the current experiment.

Time-on-task was 60 minutes in all study conditions. All students were instructed to first read the whole text. This provided a context for the main topic: sampling distribution. The experimental manipulation focused only on that part of the text (a bit longer than one page) that was about sampling distribution, and this manipulation started after all students had read the complete four pages text.

In all conditions, students were provided with a sufficient amount of paper to make notes and, depending on the condition they were in, to perform the assignments. Students in the control group were instructed to read the part on sampling distribution over and over, until the end of the 60 minutes session. In this condition, all students except one used the paper provided to make notes while reading the text, and nine of the seventeen students highlighted certain passages and/or concepts in the study text. The notes of the students were restricted largely to repeating some definitions from the text and do not indicate any self-explanation activity in this condition.

Students in the merely prompting self-explanation condition read the text part on sampling distribution once, and then received an extra document – displaying the same part of the text about sampling distribution – in nineteen separate sentences, with the instruction to self-explain every sentence, like in the experimental condition by Chi et al. (1994). Although students were prompted to self-explain every sentence, they were not asked specific questions about statistical concepts or ideas. The notes students made in this condition were restricted to self-explaining the nineteen separate sentences, none of the students made additional notes. However, seven of the sixteen students in this condition highlighted passages and/or concepts in the study text.

Students in the MPM condition read text part on sampling distribution once, then answered a total of nine propositional questions on that text, and finally performed a total of three MPM assignments after having studied the example MPM learning task and corresponding example argument displayed in Boxes 3.2. and 3.3. The notes students made in this condition focused entirely on the MPM assignments. All students managed to work out the three MPM assignments and performed each of these assignments with sufficient precision. Further, eight of the seventeen students in this condition highlighted passages and/or concepts in the study text.

Finally, all students performed a 50 minutes test that comprised two parts. The first part consisted of a total of ten multiple-choice questions and measured conceptual understanding. In each question, students had to choose the correct out of four alternatives. The questions were derived from a pool of questions about sampling distributions that had been used as exam questions in the previous years. For an example, see Box 3.4.

The second part consisted of five of the nine propositional questions students in the MPM condition had to answer and measured propositional knowledge. The questions were formulated in such a way that a short answer could be sufficient. Cognitive load was measured at the end of

the studying session as well as at the end of the testing session by means of Paas' (1992) nine-point scale.

Box 3.4.

Example of a multiple-choice question

Suppose we want to estimate the population mean by means of the sample mean. The chance that the sample mean is close to the population mean becomes larger as the sample size becomes larger. This is because:

- [A] The sample mean is equal to the expected value of the sample mean
 - [B] As the sample size increases, the variation in the sampling distribution increases as well
 - [C] As the sample size increases, the variation in the sampling distribution decreases
 - [D] Only in the case of a large sample the expected value of the mean will be equal to the population mean
-

3.2.4. Data analysis

The prior knowledge test provided one scale indicating students' prior knowledge in the form of a sum score and a second scale indicating how many misconceptions they had about the subject matter. For a randomization check, we performed multivariate analysis of variance (MANOVA), treating the two prior knowledge scales as correlated response variables ($r = -.84, p < .001$) and study condition as between-subjects factor.

For the effect of condition on cognitive load, we performed split-plot analysis of covariance (ANCOVA), using cognitive load when studying and cognitive load during the test as repeated measures (overall $r = .42, p = .003$; within conditions r ranged from .51 in the reading condition to .54 in the MPM condition) and condition as between-subjects factor. This is in line with Leppink, Broers, Imbos, Van der Vleuten, and Berger (2011b; Chapter 4 of this thesis), an article on a recently conducted experiment with a setup similar to the setup in the current experiment (only comparing slightly different treatment conditions), except that in the current analysis the two prior knowledge scales were added as quantitative covariates to our model. This was done to determine whether prior knowledge or misconceptions in prior knowledge moderates the effect of condition on cognitive load. To avoid collinearity, we used students' centered sum scores and misconception scores (instead of the original scores) in our model.

Next, we performed linear regression analysis for each of the cognitive load measures as well as for propositional knowledge and conceptual understanding. For each of these four response variables, our model included: two dummy variables for the treatment conditions (one for the unguided self-explanation condition and one for the MPM condition), students' centered sum scores and misconception scores, and the 'interactions' between each of the dummy variables and centered sum scores and misconception scores respectively. We included the latter to determine whether prior knowledge or misconceptions in prior knowledge moderates the effect of condition on the response variable in question.

Two independent raters coded the students' answers to the five propositional knowledge questions incorrect, partly correct, or correct, by comparing their answers to the descriptions derived from Moore et al. (2009). An incorrect answer was rated 0, a partly correct answer was rated 1, and a correct answer was rated 2. Initial agreement between the raters for the individual

questions was good (Cohen's $\kappa = .73$) and, for analysis purposes more importantly, the correlation between the sum scores of the two raters was high ($r = .96$) and the average difference in sum score very small. Examples of a completely correct, a partly correct, and an incorrect answer can be found in Box 3.5.

Box 3.5.

Example of how the open-ended questions were coded

Question: What is meant by the sampling distribution of a statistic?

Completely correct answer (i.e., 2 points): The sampling distribution of a statistic is the probability distribution of values of the statistic in all possible samples of the same size from the same population

Partly correct answer (i.e., 1 point): The sampling distribution of a statistic is the distribution of values taken by the statistic in more than one sample

Incorrect answer (i.e., 0 points): The sampling distribution of a statistic is the frequency distribution of (individual) scores that you find in your sample

Further, students' arguments in the MPM condition were explored to provide explanations for some of the quantitative findings.

For the multiple choice questions, incorrect choices were rated 0 and correct choices were rated 1. Given that the test consisted of five open questions and ten multiple choice questions, both propositional knowledge and conceptual understanding were measured on a scale ranging from 0 to 10.

3.3. Results

With regard to each of the four response variables (i.e., cognitive load when studying, cognitive load during the exam, propositional knowledge of statistics, and conceptual understanding of statistics) we had one hypothesis. We first present means and standard deviations for each of the response variables as well as for the two prior knowledge scales (i.e., sum score of prior knowledge and misconceptions in prior knowledge) as a function of study condition.

3.3.1. Overall analysis and randomization check

Means and standard deviations for each of the response variables and for each of the prior knowledge scales are presented in Table 3.1. MANOVA reveals that the groups did not differ significantly in prior knowledge, $F(4, 94) = 0.63, p = .64$. Further, none of the response variables displayed serious deviations from normality. In the following, we present the findings with regard to cognitive load, and subsequently, the findings with regard to propositional knowledge and conceptual understanding.

Table 3.1.
Means (and *SD*) for each of the response variables as a function of study condition

Condition	Reading (<i>n</i> = 17) <i>M</i> (<i>SD</i>)	Prompt only (<i>n</i> = 16) <i>M</i> (<i>SD</i>)	MPM (<i>n</i> = 17) <i>M</i> (<i>SD</i>)
Sum score prior knowledge (scale: 0-10)			
Ten multiple choice questions	5.76 (1.95)	5.31 (1.20)	5.76 (1.35)
Misconceptions (scale: 0-10)			
Ten multiple choice questions	2.94 (1.48)	3.56 (0.96)	3.18 (1.43)
Cognitive load (scale: 1-9)			
When studying	4.94 (1.25)	4.81 (1.33)	5.53 (1.38)
During the exam	6.24 (1.30)	5.00 (1.16)	5.00 (1.17)
Propositional knowledge score (scale: 0-10)			
Five open-ended questions	4.35 (2.64)	4.69 (2.50)	4.94 (2.25)
Conceptual understanding score (scale: 0-10)			
Ten multiple choice questions	5.82 (1.43)	5.94 (1.39)	6.41 (1.94)

3.3.2. Cognitive load

With regard to cognitive load, we hypothesized that MPM imposes less cognitive load on students than merely prompting self-explanation or not prompting self-explanation at all, in a studying phase as well as in an exam situation. The means reported in Table 3.1. are not in our expectations. Effect sizes, *F*-ratios, and *p*-values of the effects in our split-plot ANCOVA are displayed in Table 3.2. Note that in this table as well as in the tables following, ‘Misconceptions’ and ‘Sum score prior’ refer to students’ centered sum scores and misconception scores and not to the original scores.

Table 3.2.
Effect sizes, *F*-ratios, and *p*-values of the effects in split-plot ANCOVA

Effect	η^2	<i>F</i>	<i>p</i> -value
Cognitive load ^a	.11	4.89	.033
Cognitive load * Condition ^b	.32	9.82	< .001
Cognitive load * Misconceptions ^a	.01	0.46	.504
Cognitive load * Sum score prior ^a	< .01	0.09	.772
Cognitive load * Condition * Misconceptions ^b	.13	3.05	.058
Cognitive load * Condition * Sum score prior ^b	.09	1.93	.159

^a *F*(1, 41) ^b *F*(2, 41)

The significant interaction between cognitive load and study condition indicates that study condition moderates the change in cognitive load. Table 3.1. illustrates that, in the reading condition – and to a lesser extent also in the unguided self-explanation (i.e., prompt only)

condition – cognitive load was higher during the exam than in the studying phase, whereas in the guided self-explanation (i.e., MPM) condition cognitive load was lower during the exam. Table 3.3. presents mean differences in cognitive load along with 95% confidence intervals and *t*- and *p*-values per study condition.

Table 3.3.
Mean differences (*MD*) in cognitive load, 95% confidence intervals, and *t*- and *p*-values per study condition

Effect	<i>MD</i>	95% <i>CI</i>	<i>t</i>	<i>p</i> -value
Reading (control) ^a	1.30	[0.65; 1.94]	4.22	< .001
Unguided self-explanation ^b	0.19	[-0.46; 0.84]	0.61	.55
Guided self-explanation ^b	-0.53	[-1.16; 0.10]	-1.77	.10

^a *t*(16) ^b *t*(15)

In the study phase, groups did not differ significantly in average cognitive load, $F(2, 47) = 1.41$, $p = .26$. Comparing the conditions with regard to average cognitive load during the exam yielded a statistically significant effect, $F(2, 47) = 5.83$, $p = .005$. However, Table 3.1. indicates that the two self-explanation conditions do not differ at all in average cognitive load during the exam. The effect reflects a mean difference between students who self-explained (prompted or guided) and those who did not, or at least were not explicitly instructed to self-explain. Table 3.2. also indicates that students' prior knowledge, and especially misconceptions in prior knowledge, may moderate the interaction of cognitive load and study condition. Absence of statistical significance may be due to a statistical power problem. Next, we performed linear regression for the two cognitive load variables separately. Table 3.4. presents regression coefficients along with 95% confidence intervals and *t*- and *p*-values for the effects on cognitive load in the studying phase. In this table as well as in tables following, 'Unguided' and 'Guided' denote the dummy variables for the unguided self-explanation condition and the MPM condition respectively.

Table 3.4.
Regression coefficients (*B*) in cognitive load, 95% confidence intervals, and *t*- and *p*-values for the effects in the model for cognitive load in the studying phase

Effect	<i>B</i>	95% <i>CI</i>	<i>t</i>	<i>p</i> -value
Constant	4.86	[4.20; 5.52]	14.86	< .001
Unguided	-0.26	[-1.22; 0.70]	-0.55	.59
Guided	0.66	[-0.27; 1.58]	1.44	.16
Sum score prior	-0.38	[-1.07; 0.30]	-1.12	.27
Misconceptions	-0.50	[-1.40; 0.41]	-1.11	.27
Unguided * Sum score prior	1.23	[0.06; 2.39]	2.13	.040
Unguided * Misconceptions	1.89	[0.41; 3.37]	2.58	.014
Guided * Sum score prior	0.49	[-0.68; 1.65]	0.85	.40
Guided * Misconceptions	0.54	[-0.74; 1.81]	0.85	.40

Table 3.4. appears to indicate that prior knowledge and misconceptions in prior knowledge influence cognitive load when studying within the unguided self-explanation condition but not within the MPM condition. After dropping the non-significant 'Guided * Sum score prior' and 'Guided * Misconceptions' from the model, the regression coefficients and p-values of 'Unguided * Sum score prior' and 'Unguided * Misconceptions' do not change by much. A regression model within the unguided self-explanation using 'Sum score prior' and 'Misconceptions' as covariates yields similar but somewhat lower regression coefficients: For 'Sum score Prior', $B = 0.85$, $t(13) = 2.08$, $p = .058$, 95% $CI = [-0.03; 1.72]$, and for 'Misconception', $B = 1.39$, $t(13) = 2.76$, $p = .016$, 95% $CI = [0.30; 2.48]$. This model suggests that, within the unguided self-explanation condition, one point increase in the number of misconceptions in prior knowledge predicts 1.39 points (scale: 1-9) increase in cognitive load when studying. Table 3.5. presents regression coefficients along with 95% confidence intervals and t - and p -values for the effects on cognitive load during the exam.

Table 3.5.
Regression coefficients (B) in cognitive load, 95% confidence intervals, and t - and p -values for the effects in the model for cognitive load during the exam

Effect	B	95% CI	t	p -value
Constant	6.23	[5.58; 6.89]	19.25	< .001
Unguided	-1.27	[-2.22;-0.32]	-2.68	.011
Guided	-1.27	[-2.19;-0.35]	-2.80	.008
Sum score prior	-0.04	[-0.72; 0.64]	-0.13	.90
Misconceptions	-0.01	[-0.91; 0.88]	-0.03	.97
Unguided * Sum score prior	0.17	[-0.98; 1.33]	0.30	.77
Unguided * Misconceptions	0.21	[-1.25; 1.68]	0.29	.77
Guided * Sum score prior	0.33	[-0.83; 1.48]	0.57	.57
Guided * Misconceptions	0.21	[-1.05; 1.47]	0.34	.74

Table 3.5. indicates that during the exam, prior knowledge and misconceptions in prior knowledge do not influence cognitive load. This conclusion does not change after dropping the non-significant interactions. Thus, as Table 3.1. indicates, differences in cognitive load experienced in an exam situation appear to reflect a mere self-explanation effect (regardless of whether self-explanation was guided or not). The model suggests that engaging in self-explanation when studying predicts a decrease in cognitive load during the exam of 1.27 points (scale: 1-9).

3.3.3. Propositional knowledge

Table 3.6. presents regression coefficients along with 95% confidence intervals and t - and p -values for the effects in the model for propositional knowledge. From Table 3.6. follows that prior knowledge and misconceptions in prior knowledge do not influence propositional knowledge. Regression coefficients and p -values of 'Sum score prior' and 'Misconceptions' do not change significantly after dropping the non-significant interactions and the same goes for the regression coefficients and p -values of the dummy variables. Students' responses to the open-ended

questions indicate that the three study conditions are very similar in responses and typical errors and omissions. In all three conditions, students find it difficult to describe in their own words the concept of sampling distribution, and nearly all students find it difficult to explain in their own words why the sampling distribution is a probability distribution.

Table 3.6.

Regression coefficients (*B*) in cognitive load, 95% confidence intervals, and *t*- and *p*-values for the effects in the model for propositional knowledge score

Effect	<i>B</i>	95% <i>CI</i>	<i>t</i>	<i>p</i> -value
Constant	4.24	[2.98; 5.50]	6.78	< .001
Unguided	0.47	[-1.37; 2.31]	0.52	.61
Guided	0.84	[-0.93; 2.60]	0.96	.35
Sum score prior	0.48	[-0.83; 1.79]	0.74	.47
Misconceptions	-0.16	[-1.89; 1.58]	-0.18	.86
Unguided * Sum score prior	-0.30	[-2.53; 1.93]	-0.27	.79
Unguided * Misconceptions	0.25	[-2.58; 3.08]	0.18	.86
Guided * Sum score prior	-1.64	[-3.87; 0.59]	-1.49	.14
Guided * Misconceptions	-0.63	[-3.06; 1.80]	-0.52	.61

3.3.4. Conceptual understanding

Table 3.7. presents regression coefficients along with 95% confidence intervals and *t*- and *p*-values for the effects in the model for conceptual understanding.

Table 3.7.

Regression coefficients (*B*) in cognitive load, 95% confidence intervals, and *t*- and *p*-values for the effects in the model for conceptual understanding score

Effect	<i>B</i>	95% <i>CI</i>	<i>t</i>	<i>p</i> -value
Constant	5.78	[4.99; 6.57]	14.78	< .001
Unguided	0.08	[-1.07; 1.23]	0.14	.89
Guided	0.66	[-0.45; 1.77]	1.21	.24
Sum score prior	0.19	[-0.63; 1.00]	0.46	.65
Misconceptions	-0.05	[-1.13; 1.03]	-0.09	.93
Unguided * Sum score prior	0.18	[-1.21; 1.58]	0.27	.79
Unguided * Misconceptions	0.60	[-1.17; 2.37]	0.68	.50
Guided * Sum score prior	-0.75	[-2.14; 0.65]	-1.08	.29
Guided * Misconceptions	-1.09	[-2.61; 0.43]	-1.45	.16

Table 3.7. suggests that prior knowledge and misconceptions in prior knowledge do not influence the development of conceptual understanding. Dropping the non-significant 'Unguided * Sum score prior', 'Unguided * Misconceptions', and 'Unguided' does not change much in the

regression coefficients and p -values reported in Table 3.7., only the regression 'Guided * Misconceptions' increases, $B = -1.33$, $t(44) = -2.04$, $p = .047$, $95\% CI = [-2.65; -0.19]$. This appears to indicate that students who have more misconceptions in their prior knowledge profit less from the MPM approach than students who have fewer misconceptions. Subsequent regression analysis per study condition and treating students' (centered) misconception scores as covariate reveals a non-significantly negative regression coefficient in the reading condition ($B = -0.26$, $t(15) = -1.10$, $p = .29$, $95\% CI = [-0.77; 0.25]$), a non-significantly positive regression coefficient in the unguided self-explanation condition ($B = 0.18$, $t(14) = 0.48$, $p = .64$, $95\% CI = [-0.64; 1.00]$) and a statistically significant negative regression coefficient in the MPM condition ($B = -0.68$, $t(15) = -2.26$, $p = .039$, $95\% CI = [-1.33; -0.38]$). This suggests that, within the MPM condition, one point decrease in misconceptions in prior knowledge predicts an increase in conceptual understanding of 0.68 points (scale: 0-10).

When creating two misconception groups based on the median split, regression analysis using 'Guided' (i.e., the MPM dummy) as only covariate reveals that among students who have relatively many misconceptions in their prior knowledge MPM has a non-significantly negative effect on conceptual understanding, $B = -0.38$, $t(21) = -0.74$, $p = .47$, $95\% CI = [-1.46; 0.70]$, whereas among students who have relative few misconceptions MPM has a significantly positive effect on conceptual understanding, $B = 1.63$, $t(25) = 2.34$, $p = .028$, $95\% CI = [0.20; 3.06]$. These findings suggests that MPM has no (or: a rather negative than positive) effect on conceptual understanding among students who still have quite some misconceptions in their prior knowledge, but that among students who have become aware of most of their misconceptions working in MPM format predicts 1.63 points (scale: 0-10) more conceptual understanding than when not working in MPM format.

To examine the potential influence of misconceptions in prior knowledge on the development of conceptual understanding more in depth, we explored students' MPM arguments that resulted from the three MPM learning assignments. These arguments provide some explanations for why within the MPM condition, misconception awareness appears an important condition for the development of conceptual understanding.

Firstly, misconceptions with regard to randomness make it difficult for students to reason about random variables. Of the six students (in the MPM condition) who demonstrated misconceptions about randomness in the prior knowledge test, five students could not understand after having studied the text and having performed an MPM task on the topic that the sample mean in a random sample is a random variable. Although they realized that individual scores in a random sample are derived at random, they reasoned that the sample mean results from multiple individual scores and therefore cannot be considered a random variable. A quote from one of these students:

"If you draw one person from a population, you could call that a chance experiment and random because the value of this person is determined by chance. But when drawing more than one person and computing the average, you know this average is somewhere around the population mean, so you cannot call that random anymore."

The eleven students who did not demonstrate any misconceptions about randomness, on the other hand, all managed to create an argument that demonstrates their understanding of the

sample mean as a random variable. Secondly, four of the five students who did not understand the sample mean as a random variable were unable to separate the sample mean and the expected value of the sample mean, and three of them reasoned that ‘especially in a sufficiently large sample’, the sample mean is supposed to be (exactly) equal to the population mean. However, three of the twelve students who demonstrated their understanding of the sample mean as a random variable displayed the same reasoning. A closer look at the arguments of the latter three students reveals that they also had the misconception that the sampling distribution is the distribution of sample scores. A quote from one of these students:

“The 40 blood pressure values [*in a random sample of $N = 40$*] are random results and together they form the sampling distribution, so the sample mean equals 10 [*equal to the population mean*].”

Thus, students may not be able to understand the difference between sample mean and expected value of the sample mean for (at least) two different reasons; either they do not understand that the sample mean in a random sample is a random variable, or they do not understand the difference between the distribution of sample scores and the sampling distribution.

Thirdly, six of the nine students in the MPM condition who had more than the average number of misconceptions in the prior knowledge test (i.e., four or more misconceptions) did not manage to construct an argument in the last of the three MPM assignments, even after repeated attempts, and even though some of them managed to provide correct answers to the underlying open-ended questions in the assignment. Among the eight students who had less than the average number of misconceptions, only one student did not manage to create an argument. It is possible that these students would have been able to construct an argument in the task at hand, had they had more time (e.g., a few days), but this possibility could not be tested in the current experimental setup.

Taken the quantitative and qualitative findings together, it appears clear that misconception awareness is a necessary condition for MPM to have beneficial effects on conceptual understanding. However, once this condition of misconception awareness is met, MPM appears to have serious potential.

3.4. Discussion

Four hypotheses were tested. Three of the four hypotheses could not be confirmed, and with regard to our hypothesis concerning cognitive load during the exam, the decreased cognitive load during the exam was a mere self-explanation effect. In the remainder of this article, we discuss the findings reported as well as some limitations of the current experiment, and we suggest some implications for teaching practice and further research.

3.4.1. The effect of guiding self-explanation on cognitive load

We hypothesized that studying in MPM format would decrease cognitive load when studying as well as in exam situations. Table 3.1. indicates that students in the MPM condition on average experienced slightly more cognitive load than students in the other studying conditions, and that during the exam the MPM condition and the unguided self-explanation condition did not differ at all in average cognitive load. When studying, prior knowledge and misconceptions in prior

knowledge influence cognitive load within the unguided self-explanation condition but not within the MPM condition, and engaging in self-explanation – be it guided or prompted – leads to a decrease in cognitive load during the exam.

The finding that prior knowledge ($B = 0.85$) and misconceptions in prior knowledge can influence cognitive load considerably ($B = 1.39$) when self-explaining in the study phase may be explained as follows. The process of self-explanation requires students to activate their prior knowledge. When instructing students to self-explain sentences in the text, sentences that comprise one or more concepts, students appeal to their (prior) knowledge structures available in long-term memory and to the definitions they have about the concept(s) in the sentence. It is possible that a concept in the sentence is explained in a way that is slightly different from how the student defined, perhaps correctly, the concept until that point. If that is the case, the student may try to adjust his or her knowledge structure in such a way that it is more in line with the explanation in the text, and this imposes additional cognitive load on students. This holds a fortiori when the student's knowledge structure holds includes erroneous definitions or misconceptions.

It appears that studying in MPM format imposes rather more than less cognitive load on students when studying, when compared to unguided self-explanation, and that this effect does not depend on prior knowledge or misconceptions in prior knowledge within the range measured in the current study. Unfortunately, we cannot determine whether this additional load is intrinsic load, germane load, or extraneous load, for Paas' (1992) scale only provides an estimate of overall cognitive load.

3.4.2. The effect of guiding self-explanation on cognitive load

It appears that differences in propositional knowledge cannot be explained clearly by means of self-explanation or differences in prior knowledge. Even students' responses to the open-ended questions do not provide straightforward explanations for differences in propositional knowledge in terms of guided or unguided self-explanation. It is possible that some of the limitations of our study provide an explanation for this finding. Although the finding that differences in propositional knowledge cannot be explained clearly by means of self-explanation was found in a recent study by Leppink and colleagues (2011b; Chapter 4 of this thesis) as well, the latter study does indicate that prior knowledge facilitates the development of propositional knowledge.

With regard to conceptual understanding, the findings appear to indicate that within the MPM condition, misconceptions in prior knowledge have a negative effect. Students in the MPM condition scored non-significantly higher on both propositional knowledge and conceptual understanding. Moreover, the findings suggest that MPM has beneficial effect on the development of conceptual understanding only among students who have relatively few misconceptions in their prior knowledge. However, the regression coefficient of 1.63 (scale: 0-10) in this group suggests that for this group of students MPM can affect conceptual understanding greatly.

The finding that students who have relatively many misconceptions in their prior knowledge do not learn from a (guided) self-explanation approach is in line with the finding by Berthold and Renkl (2009) that self-explanation prompts do not only foster conceptual understanding but also incorrect elaborations for those who still have misconceptions about the subject matter. Within the context of basic inferential statistics, the concepts of randomness and sampling distribution

appear important concepts that students should understand before they are instructed to self-explain on these and related concepts (e.g., expected value of the sample mean). From the cognitive load perspective, an explanation for this finding is that students who have partly incorrect prior knowledge experience an extra high intrinsic load, when confronted with a learning task in which they are guided to elaborate their knowledge (Kalyuga & Hanham, 2010). In this case of high intrinsic load, any additional cognitive activities induced by guiding self-explanation may take cognitive load to the limits of working memory and lead to cognitive overload.

3.4.3. *Limitations*

The current study has a few limitations. To begin with, the open-ended questions in the exam were repeats for the MPM group. This may have put this group at an advantage on the propositional knowledge variable. Therefore, the slight mean advantage of the MPM students on the exam could very well be explained by a potentially unfair advantage. On the other hand, studying questions referring to propositions are part of the MPM procedure. This discussion could be solved by future studies using slightly different open questions in the exam.

A second limitation of the current experiment is that our conclusions with regard to cognitive load are based on Paas' (1992) scale. A pitfall of this technique is that it provides information only with regard to the total cognitive load, which has been assumed to be the sum of the three types of cognitive load. Therefore, some may criticize our interpretation by stating that the item we used to measure cognitive load does not allow us to draw straightforward conclusions with regard to potential self-explanation and prior knowledge effects on cognitive load. However, as Beckman (2010) concluded although a distinction is made between intrinsic load, extraneous load, and germane load, valid methods to measure these types of load are lacking. Measuring cognitive load by means of Paas' (1992) scale provides an indication of overall cognitive load but not of how much that cognitive load is intrinsic or extraneous or germane. We hope that future studies provide the measures that are still lacking, so that we can test hypotheses with regard to the amount of extraneous load and other load for different instructional methods.

A third limitation of our experiment is that the results are based on small sample sizes and depend upon the actual MPM assignments created for this study. Different hypotheses based on more fundamental concepts with more basic questions might lead the students to a better or worse understanding than the actual MPM assignments. With regard to the sample size, as the current study was the first study that tested MPM in an experimental setup, and the previous studies (Broers & Imbos, 2005; Broers et al., 2005) compared MPM with different conditions, it was difficult to estimate effect sizes and minimum sample sizes. Future studies could use larger sample sizes.

Besides the small sample sizes, a limitation of the current study is the small size of study matter and exam. Although the study text as a whole consisted of four pages, both the study assignments and the exam focused on a bit more than one of these four pages, and all participants completed the exam within 50 minutes. It can be questioned whether the differences between conditions with regard to propositional knowledge and conceptual understanding would have been larger, if both parts of the exam had comprised more items.

Finally, since the study was not conducted within the context of a real statistics course, the students in the current study did not have a stake in their outcomes. This may have weakened

the external validity of our study, since students may not have tried as hard as students would have if they had a stake in the outcome (e.g., a grade on their results). Therefore, a fifth limitation may be that students' learning outcomes in our study were lower than learning outcomes in the educational practice.

3.4.4. Implications for teaching practice and research

The limitations notwithstanding, the main implication for statistics education from the current study appears to be that to have students develop conceptual understanding of statistics, instructors need to help their students become aware of their misconceptions before instructing or guiding them into self-explanation activities. When studying the statistics literature, attending a lecture, or performing a learning task, students are typically confronted with important concepts and core ideas. Besides confronting students with important concepts and core ideas, lectures should also focus on frequent misconceptions students have with regard to the subject matter. In the current study, we found that, within the context of basic inferential statistics, the concepts of randomness and sampling distribution are important. Instructors should make sure that students understand these concepts before having students study and self-explain related concepts like the expected value of the sample mean. The statistics knowledge domain is a knowledge domain that is characterized by abstract concepts that are frequently built on each other. Hence, misconception awareness is a necessary condition for the development of conceptual understanding. Once the condition of misconception awareness is met, students' self-explanations and argumentations in their subsequent studying session(s) will be based on correct knowledge rather than on misconceptions. Students could then be provided with a series of MPM learning assignments on the subject matter, as the current experiment has demonstrated that once students have replaced (most of) their misconceptions by correct knowledge, MPM can help them improve their conceptual understanding considerably.

Besides that future studies should use larger studying phase(s) and exam, larger sample sizes, and different MPM assignments – perhaps for different topics within the statistics knowledge domain – an implication for future research is to focus on how and why MPM enhances conceptual understanding in certain conditions. Although the think-aloud study by Leppink and colleagues (2011a; Chapter 2 of this thesis) helped us gain insight into factors affecting the potential of MPM (e.g., propositional knowledge is a necessary but not sufficient condition for conceptual understanding, and more proficient students profit more from the MPM approach than less proficient students), we still do not know to what extent repeated practice can enhance conceptual understanding. Instead of having students work on a limited number of MPM assignments just once, future studies could focus on the benefits of having students work on a larger series of related MPM assignments, for different but related topics, and over a longer period of time. Such a practice would be more in line with a real statistics course, which typically lasts a few weeks or longer and covers multiple topics. Moreover, in this line of research, some studies could focus on the development and validation of MPM assignments for different statistics proficiency levels. It is clear that novice students cannot learn that much from assignments that are too difficult for them, and that more experienced students are not likely to learn from assignments that are too easy given their prior knowledge level. And as students grow, they need more complex assignments.

Other studies could focus on the potential of MPM in interactive or collaborative learning environments (e.g., studying in pairs, the use of MPM in classroom or lecture settings), and how prior knowledge moderates the potential of MPM in such settings. Although some studies suggest that explaining to each other, in small groups of students, yields better learning outcomes than self-explanation (e.g., Kramarski & Dudai, 2009), other studies do not confirm such a finding (Hausmann et al., 2008; Moreno, 2009), and one of the reasons for these mixed results in this comparison is that none of the studies considered potential prior knowledge effects. Since for MPM at the level of the individual student, prior knowledge is a variable to be considered (Leppink et al., 2011a, 2011b; Chapters 2 and 4 of this thesis), it would be interesting to examine to what extent prior knowledge influences learning outcomes when MPM is applied in interactive or collaborative learning settings.

A final implication for further research is that one should consider conducting it within the context of a real statistics course. In such a context, students may be more motivated to participate than outside such a context, since they have a direct stake in their outcome (e.g., an exam to be taken at the end of the course). This may lead to different learning outcomes and different effect sizes of treatments than found in the current experiment.

Chapter 4

The expertise reversal effect (I): self-explanation

Published as

Leppink, J., Broers, N. J., Imbos, Tj., Van der Vleuten, C. P. M., & Berger, M. P. F. (2011). Self-explanation in the domain of statistics: an expertise reversal effect, *Higher Education*, DOI 10.1007/s10734-011-9476-1 [open online access]

4.1. Introduction

The statistics knowledge domain is known for its abstract and cumulative nature. Although students usually develop knowledge of statistical principles and definitions (i.e., propositional knowledge, Broers, 2002) they frequently lack the ability to structure their knowledge (i.e., conceptual understanding, Broers, 2009). Knowledge elaboration can help students develop the latter ability (Kalyuga, 2009).

4.1.1. Knowledge elaboration

Knowledge elaboration is using prior knowledge to structure and integrate new information. Self-explanation (Atkinson et al., 2003; Berthold & Renkl, 2009) and argumentation (Fischer, 2002; Knipfer et al., 2009) are well-known knowledge elaboration processes. However, these processes impose cognitive load on students (Van Merriënboer & Sweller, 2005). Working memory is limited in capacity (Miller, 1956) as well as in duration (Miyake & Shah, 1999). When consciously processing new elements of information, which in complex knowledge domains like statistics are often interrelated, working memory can be overloaded (Kalyuga, 2009). Cognitive load theory assumes that the available knowledge structures in long-term memory (i.e., prior knowledge) are essential for preventing working memory overload and for guiding cognitive processes when learning (Van Merriënboer & Sweller, 2005). Most people can retain seven plus or minus two chunks of information in their working memory (Miller, 1956). What is considered a chunk of information depends on the students' prior knowledge or available knowledge structures in long-term memory. The size of a chunk is likely to increase as students' prior knowledge increases. Cognitive load imposed on students should therefore be in accordance with their prior knowledge.

Cognitive load consists of three types of load that are assumed to be additive: intrinsic load, germane load, and extraneous load. Intrinsic load depends on task complexity and students' prior knowledge of the subject. This type of load should be manipulated in instructional design by selecting learning tasks that match students' prior knowledge (Kalyuga, 2009). Germane load arises from instructional features that stimulate cognitive processes that are beneficial for learning, whereas all instructional features not directly beneficial for learning impose extraneous load on students. As the intrinsic load imposed on students when studying statistics is usually high, extraneous load should be minimized to avoid cognitive overload (Kalyuga & Hanham, 2010). By minimizing extraneous load and matching intrinsic load to the students' prior knowledge, students can engage in knowledge elaboration processes like self-explanation and argumentation, processes that impose germane load on students.

Although the process of self-explanation is time consuming, if students are given enough time, learning by self-explanation is – in terms of learning outcomes and cognitive load imposed on students – more effective than observational learning, inquiry learning, and hypermedia learning (Eysink et al., 2009). That is, self-explanation can enhance germane load activities more than the other forms of learning. Further, different studies have shown that self-explanation enhanced by prompting is more effective than spontaneous self-explanation (Atkinson et al., 2003; Chi et al., 1994). Moreover, it appears that guiding or assisting self-explanation prompts by means of open-ended questions is more effective than merely prompting self-explanation (Berthold et al., 2009). An explanation for the latter is that when students are not guided into self-explanation, they need to search the relevant subject matter themselves before they can self-explain, and this can easily lead to disorientation and therefore less efficient learning (Eysink et al., 2009). Since students guided into self-explanation do not need to search for the relevant subject matter themselves, they have more time and capacity available for self-explanation and argumentation.

4.1.2. The expertise reversal effect

Although self-explanation may be effective for some students, a question that arises is whether novice students have sufficient prior knowledge and argumentation skills to learn from self-explaining the subject matter (Kalyuga, 2009). Students who have insufficient prior knowledge experience an extra high intrinsic load, when confronted with a learning task in which they are guided to elaborate their knowledge (Kalyuga & Hanham, 2010). In this case of high intrinsic load, any additional cognitive activities induced by guiding self-explanation may take cognitive load to the limits of working memory and lead to cognitive overload. From this perspective, it is not surprising that previous studies on learning from worked-out examples indicate that novice students who have insufficient or partly incorrect prior knowledge learn more from studying worked-out examples (i.e., problems with a worked-out solution) than from solving problems or imagining solution steps themselves (e.g., Cooper et al., 2001; Kalyuga et al., 2001b; Lovett, 1992). An explanation for the latter is that for learning tasks with high intrinsic load, problem-solving imposes a high extraneous load for novice learners (Paas & Van Gog, 2006; Sweller et al., 1998).

People tend to solve new problems by searching for similar problems – of which the solution is known and the solution steps have been worked out – that can guide their solution of the new problems (Mayer, 1992). Worked-out examples of problems can guide students into self-explanation, but it depends on the students' prior knowledge (Kalyuga et al., 2001b) as well as on the design of the examples and the instructions in the examples whether students actually learn by doing so (Paas & Van Merriënboer, 1994; Van Merriënboer et al., 2002). Thus, considering students' prior knowledge is important, since it influences the effectiveness of ways to increase germane load activities like self-explanation (Paas & Van Gog, 2006). The learning activities that are intended to induce germane load will only do so if they are at a suitable level of difficulty for the student. With more prior knowledge, worked-out examples become redundant and problem solving becomes superior (Kalyuga et al., 2001b). When a learner is able to self-explain, instructional explanations as provided in worked-out examples impose extraneous load instead of germane load on the students (Kalyuga et al., 2003). The latter is also called the expertise reversal effect: there is an interaction between the levels of students' prior knowledge and the

effectiveness of different instructional methods, meaning that instructional methods that are effective for low prior knowledge students can lose their effectiveness and even have negative consequences for more proficient students (Kalyuga, 2005, 2006, 2007; Kalyuga et al., 2001a, 2003).

4.1.3. The current study

The current study investigated the expertise reversal effect in the domain of statistics, by comparing four experimental treatment conditions for low and for high prior knowledge students. In a first condition, students received a list of open-ended questions, with the instruction to answer these questions based on a study text they had to read. In a second condition, students received the same study text and the same list of open-ended questions followed by a couple of learning tasks. In each of the learning tasks, students had to create an argument integrating their answers to some of the open-ended questions to prove a statement to be true or false. In a third condition, students received the same study text as students in the first and second condition and the learning tasks from the second condition in the form of worked-out examples. Students in the fourth (control) condition were instructed to read the same study text (as students in the other conditions had to read) over and over again, and they did not receive any open-ended questions, worked-out examples, or instructions to create arguments or to perform other activities than reading the text.

Although at first including both an open-ended questions condition and an open-ended questions plus arguments condition may appear unnecessary, this inclusion is necessary to separate the effects of answering open-ended questions alone and the (subsequent) effects of creating arguments. Performance differences between the reading (control) condition and the open-ended questions plus arguments condition reflect effects of answering open-ended questions and effects of creating arguments. In this context, the open-ended questions condition could be regarded as a second control condition.

The treatment conditions in the current experiment are the same as used by Broers and Imbos (2005). However, in the current study students' differences in prior knowledge were taken into account to examine potential interaction effects between students' prior knowledge and the effectiveness of different instructional methods (i.e., expertise reversal effects). Further, time-on-task was kept constant in the current study. When ignoring the effect of time-on-task, we cannot determine whether (potentially) beneficial effects in terms of learning outcomes can be attributed to effective aspects of the instructional method or that it is due to the single fact that students in one group spend more time on their assignments than students in other groups. If the latter is the case, it is questionable whether in practice students will choose to work according to this rather time-consuming method. Finally, cognitive load was measured in the current study. Even if an instructional method has potentially beneficial effects, it is important to know how much effort it takes for students (i.e., how much cognitive load is imposed on them) to learn according to this method. Besides, an instructional method requiring more effort may decrease cognitive load during exam situations, if learning according to this method enables students to enhance their knowledge structures in long-term memory (Van Merriënboer & Sweller, 2005).

In line with previous studies on the expertise reversal effect we expected that low prior knowledge students learn relatively more from worked-out examples, whereas high prior

knowledge students learn relatively more from answering open-ended questions and formulating arguments. In other words, we expected that among low prior knowledge students' propositional knowledge and conceptual understanding are elevated mostly when studying worked-out examples, whereas the combination of answering open-ended questions and formulating arguments yields optimal propositional knowledge and conceptual understanding among high prior knowledge students.

In cognitive load theory, developing propositional knowledge and conceptual understanding means enhancing knowledge structures about the subject matter in long-term memory (Van Merriënboer & Sweller, 2005). Students who have enhanced knowledge structures about the subject matter in their long-term memory are likely to experience lower cognitive load than their less knowledgeable peers when confronted with the subject matter or in an exam situation. Therefore, we expected that when studying statistics, low prior knowledge students experience more cognitive load when studying than high prior knowledge students. Further, in line with previous studies on the expertise reversal effect, we expected that low prior knowledge students who learn from worked-out examples experience less cognitive load during an exam than low prior knowledge students who answer open-ended questions and formulate arguments, whereas for high prior knowledge students the trend goes in the other direction. Thus, the following three hypotheses were tested:

- (1) Low prior knowledge students experience more cognitive load when studying statistics than high prior knowledge students.
- (2) Low prior knowledge students who learn from worked-out examples experience less cognitive load during an exam than low prior knowledge students who answer open-ended questions and formulate arguments, whereas for high prior knowledge students the trend goes in the other direction.
- (3) Among low prior knowledge students propositional knowledge and conceptual understanding are elevated most when studying worked-out examples, whereas answering open-ended questions and formulating arguments yields optimal propositional knowledge and conceptual understanding among high prior knowledge students.

4.2. Method

The current study investigated the effects of four instructional methods on cognitive load, propositional knowledge, and conceptual understanding of statistics, for low prior knowledge students and for high prior knowledge students.

4.2.1. Participants and experimental design

A total of 130 first-year bachelor students in psychology and health sciences who had not yet attended any university statistics course were divided into two groups, based on the median split of their prior knowledge scores on a questionnaire measuring statistical reasoning ability. Although originally we considered prior knowledge as a continuous covariate in the model, the distribution of prior knowledge scores was did not approach a symmetric and unimodal distribution but rather showed two prior knowledge subgroups. Therefore, we chose to include prior knowledge as a factor consisting of two levels, being a low prior knowledge and high prior knowledge group. Within both groups, students were allocated at random to one of the four

possible treatment conditions: a reading-only (control) conditions, an open-ended questions condition, an open-ended questions plus arguments condition, and a worked-out examples condition. Twenty-five students had to cancel their participation due to unexpected changes in their educational timetable. As students did not know before the studying session in what condition they would participate, this drop-out was not a consequence of any experimental treatment.

4.2.2. Materials

Materials were: (1) a pretest on statistical reasoning (a subset from the Statistical Reasoning Assessment, Garfield, 2003); (2) a text of four pages on basic inferential statistics, composed by the authors of the manuscript from chapters 4-6 of Moore et al. (2009), that had been subjected to a pilot-study for assessing its difficulty level and time required to read it properly; (3) one study task per group; (4) a Dutch validated version of Paas' (1992) nine point scale for measuring cognitive load; and (5) a 50 minutes test consisting of a part measuring propositional knowledge and another part measuring conceptual understanding.

4.2.3. Procedure

After finishing the pretest on statistical reasoning and being randomly allocated to one of the experimental treatment conditions, students were presented the text on the sampling distribution of the mean and they were instructed to work for 60 minutes. All students were instructed to first read the whole text. This provided a context for the main topic: sampling distribution. The experimental manipulation focused on the part of the text (a bit longer than one page) that was about sampling distribution, and this manipulation started after students had read the complete four pages text.

Students in the control condition were instructed to read the text part on sampling distribution over and over, until the end of the 60 minutes session. Students in the open-ended questions condition read the text part on sampling distribution once, and then answered a total of nine open-ended questions on that text. Three of these questions are displayed in Box 4.1.

Box 4.1.

Example of open-ended questions

-
- [1] What is a sampling distribution of an estimator?
 - [2] What is the expected value of an estimator?
 - [3] What is meant by variation in the values of an estimator?
-

Students in the open-ended questions plus arguments condition read the text part on sampling distribution once, and then received a document displaying the same nine open-ended questions as in the open-ended questions condition followed by three learning tasks. One of these learning tasks is displayed in Box 4.2.

Box 4.2.

Example of a learning task in the open-ended questions plus arguments condition

In a population Q , variable X follows a Normal distribution and its mean is 10. We draw a random sample of size N from this population.

Hypothesis: *The sample mean equals 10.*

Questions to be processed in the argument:

- [1] What is a sampling distribution of an estimator?
 - [2] What is the expected value of an estimator?
 - [3] What is meant by variation in the values of an estimator?
-

Each learning task comprised a true/false statement and three of the nine open-ended questions (as presented in Box 4.1.), and in each learning task students were instructed to formulate an argument integrating the answers to the three open-ended questions that could prove the statement to be true or false. Finally, students in the worked-out examples condition first read the text part on sampling distribution and then study the three learning tasks from the open-ended questions plus arguments condition in the form of worked-out examples. For a worked-out example of an argument see Box 4.3.

Box 4.3.

Example of a worked-out example of an argument

There is variation in variable X in population Q . The drawn sample of size N is just one of the many possible samples of this size N we could draw from this population. An estimator like the sample mean varies over these possible samples and thus has variation [answer to question 3 in Box 4.2.]. The sampling distribution of an estimator is the probability distribution of the values of the estimator over all possible samples of the same sample size N from the same population [answer to question 1 in Box 4.2.]. Thus, the sampling distribution is about variation in the estimator [answers to questions 1 and 3 related]. The expected value of the estimator is the average value of the estimator over all possible samples of the same sample size N from the same population [answer to question 2 in Box 4.2.]. Thus, the expected value is the mean of the sampling distribution [answers to questions 1 and 2 related] and this equals the value of the parameter to be estimated, in our case the population mean. The expected value of the sample mean equals 10, but the mean that we find in our sample is not necessarily equal to 10. Conclusion: the hypothesis is false.

Thus, in all treatment conditions except the reading-only (control) condition students were confronted with open-ended questions. Students in the worked-out examples condition were given the answers to these questions, while students in the other two conditions had to answer these questions themselves. Moreover, students in the worked-out examples condition and the open-ended questions plus arguments condition learned from arguments, with the difference that students in the latter condition had to formulate these arguments themselves.

After the 60 minutes studying session, students did a 50 minutes test consisting of two parts. The first part consisted of a total of ten multiple-choice questions, typically embedded in some problem context, measuring conceptual understanding. These questions were derived from a pool of questions about the sampling distribution that had been used as exam questions in the previous years. For an example, see Box 4.4.

Box 4.4.

Example of a multiple-choice question

Suppose we want to estimate the population mean by means of the sample mean. The chance that the sample mean is close to the population mean becomes larger as the sample size becomes larger. This is because:

- [A] The sample mean is equal to the expected value of the sample mean
 - [B] As the sample size increases, the variation in the sampling distribution increases as well
 - [C] As the sample size increases, the variation in the sampling distribution decreases
 - [D] Only in the case of a large sample the expected value of the mean will be equal to the population mean
-

For each question, students had to choose the correct one out of four alternatives. The second part consisted of five out of the nine open-ended questions students in all conditions except the reading-only (control) condition had to answer. We chose these questions to measure how much propositional knowledge students had developed after the study assignment and to compare the average propositional knowledge between conditions. The questions were formulated in such a way that a short answer would be sufficient. Cognitive load was measured twice, namely at the end of the studying session and at the end of the test.

4.2.4. Data analysis

We performed analyses of variance (ANOVA) having as factors instructional method (i.e., experimental treatment condition, four levels) and prior knowledge group (i.e., high vs. low). For propositional knowledge and conceptual understanding, we performed two separate two-way ANOVAs, while for cognitive load we performed split-plot ANOVA using cognitive load when studying and cognitive load during the test as repeated measures. Two independent raters coded the open-ended questions by comparing students' short answers to Moore et al.'s (2009) formulations. An incorrect answer was rated 0, a partly correct answer was rated 1, and a completely correct answer was rated 2. Initial agreement between the raters for the individual questions was high (Cohen's $\kappa = .87$), the correlation between the sum scores of the two raters was also high ($r = .97$) and the average difference in sum score very small. Examples of a completely correct, a partly correct, and an incorrect answer can be found in Box 4.5. For the multiple choice questions, incorrect choices were rated 0 and correct choices were rated 1. Given that the test consisted of five open-ended questions and ten multiple choice questions, both propositional knowledge and conceptual understanding were measured on a scale ranging from 0 to 10.

Box 4.5.

Example of how the open-ended questions were coded

Question: What is meant by the sampling distribution of an estimator?

Completely correct answer (i.e., 2 points): The sampling distribution of an estimator is the probability distribution of values of the estimator in all possible samples of the same size from the same population

Partly correct answer (i.e., 1 point): The sampling distribution of an estimator is the distribution of values taken by the estimator in more than one sample

Incorrect answer (i.e., 0 points): The sampling distribution of an estimator is the frequency distribution of (individual) scores that you find in your sample

4.3. Results

We first present the results with regard to cognitive load, and next, the findings with regard to propositional knowledge and conceptual understanding.

4.3.1. Cognitive load

First of all, we expected a prior knowledge effect on cognitive load when studying statistics (i.e., first hypothesis). Our second expectation with regard to cognitive load (i.e., second hypothesis) was that low prior knowledge students who learn from worked-out examples experience less cognitive load during an exam than low prior knowledge students who answer open-ended questions and formulate arguments, whereas for high prior knowledge students the trend goes in the other direction. Means and standard deviations (*SD*) of cognitive load when studying and cognitive load during the exam are presented in Table 4.1.

Table 4.1.
Means (and *SD*) and sample sizes (*n*) of cognitive load

Group	Low prior knowledge	High prior knowledge
Cognitive load when studying (scale: 1-9)		
Reading-only	4.83 (1.19) <i>n</i> = 12	5.25 (1.96) <i>n</i> = 12
Questions	5.64 (1.99) <i>n</i> = 14	5.23 (2.13) <i>n</i> = 13
Questions and arguments	6.21 (1.81) <i>n</i> = 14	5.93 (1.59) <i>n</i> = 14
Worked-out examples	5.92 (1.19) <i>n</i> = 13	6.62 (1.04) <i>n</i> = 13
Cognitive load during the exam (scale: 1-9)		
Reading-only	6.25 (0.97) <i>n</i> = 12	6.17 (1.19) <i>n</i> = 12
Questions	6.00 (1.36) <i>n</i> = 14	5.62 (2.02) <i>n</i> = 13
Questions and arguments	6.14 (0.95) <i>n</i> = 14	5.57 (1.70) <i>n</i> = 14
Worked-out examples	6.23 (1.17) <i>n</i> = 13	6.38 (1.50) <i>n</i> = 13

Split-plot ANOVA using instructional method and prior knowledge group as factors and treating cognitive load when studying and cognitive load during the test as repeated measures revealed a non-significant three-way interaction, $F(3, 97) = 0.175, p > .90 (\eta^2 = .005)$ as well as a non-significant two-way interaction between cognitive load and prior knowledge group, $F(1, 97) = 1.087, p > .30 (\eta^2 = .011)$. However, the interaction between cognitive load and instructional method was statistically significant, $F(3, 101) = 3.690, p < .05 (\eta^2 = .099)$. Subsequently, we performed separate two-way ANOVAs for each of the cognitive load measurements, using instructional method and prior knowledge group as factors.

With regard to cognitive load when studying, the interaction effect was very small and not statistically significant, $F(3, 97) = 0.684, p > .55 (\eta^2 = .021)$. However, the main effect of instructional method was significant, $F(3, 101) = 2.991, p < .05 (\eta^2 = .082)$. Post-hoc comparisons using Tukey's correction for multiple testing revealed a significant difference between the control group and the worked-out examples group, $t(48) = 2.63, p < .05$, the average cognitive load when studying being highest in the worked-out examples condition. The main effect of prior knowledge was negligible, $F(1, 103) = 0.073, p > .75 (\eta^2 = .001)$. Thus, with regard to cognitive load when studying there appears to be a main effect of instructional method rather than a prior knowledge effect.

With regard to cognitive load during the test, the interaction effect was very small and not statistically significant, $F(3, 97) = 0.348, p > .75 (\eta^2 = .011)$. The same holds for the main effect of instructional method, $F(3, 101) = 0.837, p > .45 (\eta^2 = .024)$, as well as for the main effect of prior knowledge, $F(1, 103) = 0.707, p > .40 (\eta^2 = .007)$.

Thus, with regard to cognitive load we can conclude that although there is a medium size effect of instructional method when studying, in test situations this effect is reduced to small and virtually non-existing.

4.3.2. Propositional knowledge and conceptual understanding

We expected the effect of instructional method on propositional knowledge and conceptual understanding to be different for high prior knowledge students than for low prior knowledge students (i.e., third hypothesis). More specifically, we expected that low prior knowledge students would benefit most from studying worked-out examples, whereas high prior knowledge students would benefit most from answering open-ended questions and formulating arguments. This hypothesis was confirmed for conceptual understanding, but not for propositional knowledge. Means and standard deviations of propositional knowledge are presented in Table 4.2. The interaction effect was very small and not statistically significant, $F(3, 97) = 0.389, p > .75 (\eta^2 = .012)$. The same holds for the main effect of instructional method, $F(3, 101) = 0.609, p > .60 (\eta^2 = .018)$. The main effect of prior knowledge was statistically significant, $F(1, 103) = 4.366, p < .05 (\eta^2 = .041)$. These findings illustrate a small to moderate prior knowledge effect on propositional knowledge, independent of instructional method. The means and standard deviations for total illustrate the main effect for prior knowledge on propositional knowledge. High prior knowledge students scored on average almost one point higher than low prior knowledge students, $t(103) = 2.089, p < .05$ (Cohen's $d = 0.41$).

Table 4.2.
Means (and *SD*) and sample sizes (*n*) of propositional knowledge score

Group	Low prior knowledge	High prior knowledge
Propositional knowledge score (scale: 0-10)		
Reading-only	4.33 (2.54) <i>n</i> = 12	5.33 (2.23) <i>n</i> = 12
Questions	5.29 (2.64) <i>n</i> = 14	6.15 (2.15) <i>n</i> = 13
Questions and arguments	4.57 (1.83) <i>n</i> = 14	6.14 (1.92) <i>n</i> = 14
Worked-out examples	5.23 (2.74) <i>n</i> = 13	5.46 (2.15) <i>n</i> = 13

Although the hypothesis of differential instructional method effects for the different prior knowledge groups was not confirmed for propositional knowledge, it was confirmed for conceptual understanding. We found a significant interaction effect, $F(3, 97) = 2.762, p < .05 (\eta^2 = .079)$. Means and standard deviations of conceptual understanding are presented in Table 4.3.

Table 4.3.
Means (and *SD*) and sample sizes (*n*) of conceptual understanding score

Group	Low prior knowledge	High prior knowledge
Conceptual understanding score (scale: 0-10)		
Reading-only	4.83 (1.34) <i>n</i> = 12	5.42 (1.56) <i>n</i> = 12
Questions	5.14 (1.70) <i>n</i> = 14	5.00 (2.08) <i>n</i> = 13
Questions and arguments	4.64 (1.08) <i>n</i> = 14	6.79 (1.31) <i>n</i> = 14
Worked-out examples	6.00 (1.78) <i>n</i> = 13	5.85 (2.41) <i>n</i> = 13

Given the interaction pattern, main effects are difficult to interpret. Although two-way ANOVA yields a marginally significant main effect for prior knowledge, $F(1, 97) = 3.32, p < .10 (\eta^2 = .033)$ and also a small to medium size (although non-significant) main effect for instructional method, $F(3, 97) = 1.625, p < .20 (\eta^2 = .048)$, the interaction pattern does not allow a straightforward interpretation of these effects. For example, in the open-ended questions plus arguments condition, high prior knowledge students scored on average more than two points higher than low prior knowledge students, whereas in the worked-out examples condition as well as in the open-ended questions (only) condition high prior knowledge students performed rather worse than low prior knowledge students. Subsequent one-way ANOVAs per prior knowledge group are non-significant, but this is most likely due to a lack of statistical power, since the effect sizes are rather large: for high prior knowledge students, $F(3, 48) = 2.219, p < .10 (\eta^2 = .122)$, and for low prior knowledge students, $F(3, 49) = 2.097, p < .15 (\eta^2 = .114)$.

The findings with regard to conceptual understanding illustrate an expertise reversal effect. Low prior knowledge students learn most from studying worked-out examples, while the combination of open-ended questions and arguments is least effective, $t(25) = 2.415, p < .023$ (Cohen's $d = 0.96$). High prior knowledge students, though, learn most from the combination of

open-ended questions and arguments, while answering open-ended questions only is least effective, $t(25) = 2.688, p < .05$ (Cohen's $d = 1.06$), studying worked-out examples not yielding significantly worse learning outcomes than the combination of open-ended questions and arguments, $t(25) = 1.271, p > .20$ (Cohen's $d = 0.51$). The non-significant difference between the open-ended questions plus arguments condition and the worked-out examples condition may be due to a statistical power problem, since the difference illustrates a medium size effect.

4.4. Discussion

The results indicate a dominant prior knowledge effect on propositional knowledge, an effect of instructional method on cognitive load when studying, and an effect of instructional method on conceptual understanding that depends on students' prior knowledge. In the remainder of this article, we discuss these findings as well as some limitations of the current experiment and we suggest some implications for the teaching practice and for future research.

4.4.1. The effect of instructional method on cognitive load

At first, the findings with regard to cognitive load do not appear to illustrate an expertise reversal effect. However, when evaluating the findings with regard to cognitive load in the light of propositional knowledge and conceptual understanding, such an effect appears to become visible.

During the studying session, on average cognitive load was highest in the worked-out examples condition followed by the open-ended questions plus arguments condition (see Table 4.1.). Further, in the low prior knowledge group cognitive load was highest in the open-ended questions plus arguments condition, whereas in the high prior knowledge group cognitive load was highest when studying worked-out examples. The latter contrast is more or less the opposite from the contrast we found for conceptual understanding (see Table 4.3.): high prior knowledge students performed best when answering open-ended questions and formulating arguments, whereas low prior knowledge students performed best when studying worked-out examples and worst when answering open-ended questions and formulating arguments. We interpret this expertise reversal effect as follows. On the one hand, for low prior knowledge students, studying worked-out examples imposes a high germane load, whereas formulating arguments integrating answers to open-ended questions rather imposes a high extraneous load. As a consequence, in this group of students worked-out examples enhance conceptual understanding much more than formulating arguments. On the other hand, for high prior knowledge students, it is the process of formulating arguments that imposes a high germane load, whereas worked-out examples rather impose a higher extraneous load. As a consequence, in this group of students, formulating arguments enhances conceptual understanding more than worked-out examples.

Some may criticize our interpretation, since Paas' (1992) scale provides an indication of overall cognitive load and we have not included items to measure the different types of cognitive load. Although we are aware that such items have been proposed (e.g., Eysink et al., 2009), as these and other items have not yet been subjected to proper validation research (Beckmann, 2010) we were reluctant to use them. We hope that future studies can provide validated items for the different types of load, for we wish to test hypotheses with regard to the amount of extraneous load and other load for different instructional methods. As long as no validated items

for the different types of cognitive load are at hand, we prefer interpretations based on total cognitive load.

Efficiency scores combining total cognitive load scores with results on achievement tests are commonly carried out in cognitive load theory sourced experiments (Paas & Van Merriënboer, 1993; Van Gog & Paas, 2008). These measures, however, do not indicate how knowledge differences between students but rather how much effort students needed to develop the knowledge they have developed. Further, since subjective rating scales are often not very sensitive to variations in actual cognitive load, performance scores usually provide the best available evidence for expertise reversal effects.

4.4.2. The effect of prior knowledge on propositional knowledge of statistics

Although in the worked-out examples condition the effect is not as strong as in the other conditions, on average one can conclude that low prior knowledge students find it more difficult to develop propositional knowledge (see Table 4.2.) than high prior knowledge students. As with regard to propositional knowledge the interaction between instructional method and prior knowledge group is very small and non-significant, there is no clear indication for an expertise reversal effect for this type of knowledge. Propositional knowledge is knowledge of more or less isolated statistical concepts and ideas and it appears that, regardless of the instructional method, for low prior knowledge students it is more difficult to develop this type of knowledge.

If there is an expertise reversal effect with regard to propositional knowledge that we did not find, this may be due to the small size of study matter and test. Although the study text as a whole consisted of four pages, both the study assignments and the exam focused on a bit more than one of these four pages, and all participants completed the exam within 50 minutes. Differences between conditions as well as the interaction effect between study condition (i.e., instructional method) and prior knowledge group might have been larger, if the exam had comprised more items.

4.4.3. Conceptual understanding of statistics: an expertise reversal effect

We already evaluated the expertise reversal effect with regard to conceptual understanding in the context of our findings with regard to cognitive load. When compared to worked-out examples, we found positive argumentation effects on conceptual understanding for high prior knowledge students, whereas for low prior knowledge students it affects conceptual understanding negatively. When comparing the open-ended questions condition and the open-ended questions plus arguments condition on conceptual understanding (see Table 4.3.), a similar pattern becomes visible. For low prior knowledge students it is better to spend more time on answering open-ended questions than spending part of their time on formulating arguments. This conclusion appears to hold for propositional knowledge as well (see Table 4.2.). Thus, for this group of students, argumentation based on their answers to open-ended questions negatively affects propositional knowledge and conceptual understanding, and can even lead to worse conceptual understanding than when merely reading the study text (see Table 4.3.). On the other hand, high prior knowledge students perform worst on conceptual understanding – even worse than reading the study text without seeing any open-ended questions – when answering open-ended questions without subsequent knowledge elaboration by means of argumentation.

The finding that knowledge elaboration by means of formulating arguments in which answers to open-ended questions are integrated only works for high prior knowledge students appears easy to explain: even if you have strong argumentation skills, if you lack (prior) knowledge, how can you draw valid conclusions? Argumentation based on flawed or incomplete knowledge does not contribute to learning and is therefore likely to impose extraneous load on students. Only students who have sufficient prior knowledge are likely to engage in germane load activities and to develop a better understanding by means of argumentation.

Some people may argue that another limitation of the current experiment is that students in the reading-only condition did not receive the open-ended questions in the studying session, and that therefore these students had an unfair disadvantage during the test when compared to the other three experimental treatments. However, by demonstrating that low prior knowledge students who have the open-ended questions in the studying session on average perform worse than students who do not see the open-ended questions, we think that this is not a serious limitation of our experiment.

4.4.4. Implications for the teaching practice and for new research

The limitations with regard to cognitive load, the small size of study matter and test, and the open-ended questions issue notwithstanding, we have a clear implication for the teaching practice as well as for future research.

Confronting students with worked-out examples of arguments (see Box 4.3. for an example) appears to be a good initiative for teachers introducing the subject matter to be learned, and for elaborating on students' limited prior knowledge. However, once students have sufficient (prior) knowledge, knowledge elaboration by means of argumentation (see Box 4.2. for an example) imposes germane load on students, while reducing overall load (see Table 4.1.). Thus, worked-out examples being a good starting initiative, once students have more (prior) knowledge it is time for them to actively self-explain the statistical concepts and ideas and integrate them into arguments to solve more complex problems. Further, as students step by step develop more conceptual understanding, they gradually need less instructional guidance when solving problems of a certain complexity level, or they can be guided into more complex problems. Given a certain subject matter and complexity level, educational practice could strive for decreasing guidance as the course proceeds:

- *Guided problem-solving* (e.g., Box 4.3.): students are confronted with worked-out examples;
- *Semi-guided or self-guided problem-solving* (e.g., Box 4.2.): students need to solve problems by bearing in mind a limited set of rules, propositions, or premises (e.g., integrating an indicated number of propositions into an argument); and
- *Unguided problem-solving*: this situation is closer to real-life practice of statistics, in which statisticians, researchers, or other experts dealing with statistics, need to select the relevant rules, propositions, or premises themselves, and apply their knowledge and understanding autonomously, without guidance.

Depending on the course aims, a combination of the self-guided and unguided problem-solving approach could be applied in the exam (e.g., two different types of exam problems). When applying the unguided approach in an exam following a course applying the guided and

self-guided approach, an interesting question is how students solve the problems in an exam, more specifically: will students still solve the problems by integrating relevant propositions into arguments? If yes, they demonstrate conceptual understanding, and that they can apply (some of) their knowledge and communicate with others in a way that makes sense. Alevan and Koedinger (2002) demonstrated that the more unguided approach can be effective in geometry. Future studies could examine the effectiveness for the unguided approach for students varying in prior knowledge. The current experiment demonstrates that low prior knowledge students profit most from studying worked-out examples, whereas high prior knowledge students profit most from formulating arguments. Perhaps future studies will demonstrate that at a next level, students learn more from the more unguided approach.

Chapter 5

The expertise reversal effect (I): self-explanation

Published as

Leppink, J., Broers, N. J., Imbos, Tj., Van der Vleuten, C. P. M., & Berger, M. P. F. (2012). Prior knowledge moderates instructional effects on conceptual understanding of statistics, *Educational Research and Evaluation*, 18, 37-51

5.1. Introduction

Statistics is known to be a difficult subject. Although students usually develop knowledge of definitions, statistical ideas, and statements (i.e., propositional knowledge), many students find it difficult to interrelate and structure their knowledge (i.e., conceptual understanding; Broers, 2009). The current article addresses the practical problem of how self-explanation, explanation to others, and extra instruction in learning tasks may help students who failed their statistics exam find motivation to resume their study of statistics and improve their knowledge and understanding of statistics.

5.1.1. Self-explanation and the added value of explanation to peers

Self-explanation is a form of constructive learning that has been proven to enhance knowledge and understanding of learning material more than merely reading the learning material (i.e., passive learning), highlighting passages of text or repeating text sentences verbatim (i.e., active learning; Fonseca & Chi, 2011). Moreover, when compared to other constructive learning strategies (e.g., concept mapping, drawing diagrams, generative summarizing), self-explanation appears to give superior learning outcomes. Although in the latter comparison effect sizes vary from large (Roscoe & Chi, 2008) to very small (King, 1992), beneficial self-explanation effects on learning and problem-solving have been demonstrated across a variety of domains and settings (Atkinson et al., 2003). The two most successful cognitive mechanisms of self-explanations appear to be filling knowledge gaps (Chi, 2000) and constructing knowledge networks (Novak, 2002).

Constructive learning strategies, like self-explanation, require students to produce an output comprising information that goes beyond that provided in the original study material. Although constructive learning strategies generally yield better learning outcomes than passive or active learning strategies, some argue that receiving additional information in the form of feedback, elaborations, or questions, are crucial components of interactive learning situations that may go beyond the beneficial effects of self-explanation and other constructive learning strategies, even if this information is given by peers. For example, Kramarski and Dudai (2009) found that explaining to each other, in groups of four students, enhanced mathematical understanding more than engaging in self-explanation. Moreno (2009), however, found no differences between interactive learning and self-explanation in terms of learning outcomes, and in a study by Hausmann et al. (2008), the findings were mixed. In the latter study, students worked on a number of physics problems either alone (self-explanation) or in pairs (joint explanation). The pairs responded faster to questions, solved problems more quickly, gave more correct responses,

and needed less hints. However, the two conditions (i.e., explaining in pairs or explaining to oneself) did not differ significantly on a shallow definitions post-test, and on a deeper learning outcomes test, the self-explanation group performed significantly better than the joint explanation group.

One of the reasons for the somewhat mixed results with regard to the comparison of self-explanation and interactive learning may be that the studies reported did not take into account potential prior knowledge effects. A recent study by Leppink et al. (2011a; Chapter 2 of this thesis) illustrates that whether students learn from self-explanation depends on students' prior knowledge about the subject matter.

5.1.2. Prior knowledge and additional information in learning tasks

Learning imposes cognitive load on students (Van Merriënboer & Sweller, 2005). Working memory is limited in capacity as well as in duration (Miyake & Shah, 1999). When consciously processing new elements of information, which in complex knowledge domains like statistics are often interrelated, working memory can be overloaded (Kalyuga, 2009). Cognitive load theory assumes that the available knowledge structures in long-term memory (i.e., prior knowledge) are essential for preventing working memory overload as well as for guiding cognitive processes when learning (Van Merriënboer & Sweller, 2005). Cognitive load imposed on students should therefore be in accordance with their prior knowledge.

Cognitive load consists of three types of load that are assumed to be additive: intrinsic load, germane load, and extraneous load. Intrinsic load depends on task complexity and students' prior knowledge about the subject matter. This type of load should be manipulated in instructional design by selecting learning tasks that match students' prior knowledge (Kalyuga, 2009). Germane load arises from instructional features that stimulate cognitive processes beneficial for learning, whereas all instructional features not directly beneficial for learning impose extraneous load on students. As the intrinsic load imposed on students when studying statistics is usually high, extraneous load should be minimized to avoid cognitive overload (Kalyuga & Hanham, 2010). By minimizing extraneous load and matching intrinsic load to the students' prior knowledge, students can engage in self-explanation, or joint explanation, and this imposes germane load on students.

Although processes like self-explanation or explanation to peers may be effective for some students, the question is whether novice students have sufficient prior knowledge and argumentation skills to learn from explaining the subject matter (Kalyuga, 2009). Students who have insufficient prior knowledge experience an extra high intrinsic load, when confronted with a learning task in which they are guided to elaborate their (prior) knowledge (Kalyuga & Hanham, 2010). In this case of high intrinsic load, any additional cognitive activities induced by guiding self-explanation or joint explanation may take cognitive load to the limits of working memory and lead to cognitive overload. From this perspective, it is not surprising that previous studies on learning from worked-out examples indicate that novice students who have insufficient or partly incorrect prior knowledge learn more from studying worked-out examples (i.e., problems with a worked-out solution) than from solving problems or imagining solution steps themselves (e.g., Cooper et al., 2001; Kalyuga et al., 2001b; Lovett, 1992). An explanation for the latter is that for learning tasks with high intrinsic load, problem-solving imposes a high extraneous load for novice learners (Paas & Van Gog, 2006; Sweller et al., 1998).

People tend to solve new problems by searching similar problems – of which the solution is known and the solution steps have been worked out – that can guide their solution of the new problems (Mayer, 1992). Worked-out examples of problems can guide students into self-explanation, but it depends on the students' prior knowledge (Kalyuga et al., 2001) as well as on the design and instructions (Paas & Van Merriënboer, 1994; Van Merriënboer et al., 2002) of the worked-out examples whether students actually learn by doing so. Thus, considering students' prior knowledge is important, since it influences the effectiveness of ways to increase germane load activities (Paas & Van Gog, 2006) like self-explanation or explanation to peers. The learning activities that are intended to induce germane load will only do so if they are at a suitable level of difficulty for the student. With more prior knowledge of the subject, worked-out examples become redundant and problem solving becomes superior (Kalyuga et al., 2001b), because when a learner is able to explain, instructional explanations as provided in worked-out examples or partially worked-out examples are redundant and therefore impose extraneous load instead of germane load on the students (Kalyuga et al., 2003).

Students' prior knowledge moderates the effectiveness of different instructional methods, meaning that instructional methods that are effective with low prior knowledge students lose their effectiveness and even have potentially negative consequences for more knowledgeable students (Kalyuga, 2005, 2006, 2007; Kalyuga et al., 2003). The more prior knowledge students have, the less (additional) information is needed in their learning tasks. To optimize learning for students varying in prior knowledge level, explanation to peers should be planned and structured based on students' prior knowledge, and likewise, additional information in learning tasks should be tailored to the students' prior knowledge.

5.1.3. The current experiment

The current experiment compared the effects of self-explaining and explaining in pairs a couple of learning tasks on basic inferential statistics – for low prior knowledge students and for students having more prior knowledge – on cognitive load, propositional knowledge, and conceptual understanding. In line with the study by Leppink et al. (2011a; Chapter 2 of this thesis) suggesting that the effects of self-explanation on learning outcomes depend on students' prior knowledge, we hypothesized that low prior knowledge students self-explaining the learning materials would develop less propositional knowledge and conceptual understanding than students who have more prior knowledge. Low prior knowledge students would need additional information to be able to learn from their self-explanations.

Additional information can come from the learning task itself, from a peer doing the same learning task, or from a combination of both. Since two students usually know more than one student (two students having no knowledge at all being an exception), explaining a learning task in pairs is an example of receiving and using additional information from your peer. Assuming that additional information adds up to the student's prior knowledge (Johnson et al., 2007), we hypothesized that low prior knowledge students who receive additional information – either from the learning task itself or from a peer – develop more propositional knowledge and conceptual understanding than low prior knowledge students who do not receive additional information. For students who have more prior knowledge, however, receiving additional information might be redundant and therefore not lead to enhanced propositional knowledge or conceptual understanding.

Thus, we hypothesized that receiving additional information – be it from the learning task itself or from a peer – would not enhance propositional knowledge and conceptual understanding among students who already have more prior knowledge. Testing this hypothesis and the previously mentioned hypotheses requires three factors: (1) students self-explaining vs. students explaining to each other (i.e., in pairs), (2) learning tasks comprising additional information vs. learning tasks not comprising additional information, and (3) low prior knowledge students vs. students who have (slightly) more prior knowledge. Although one might consider the prior knowledge variable as a continuous covariate in the model, the distribution of prior knowledge scores in our experiment suffered from restriction of range and strongly deviated from a symmetric and unimodal distribution. Therefore, we chose to include prior knowledge as a factor consisting of the aforementioned two levels. Further, one may ask why it is interesting to split up these students instead of pairing them off (i.e., each pair consisting of one low prior knowledge student and one student having more prior knowledge) in the experimental conditions. Although we considered this option, in such a design the more proficient student explaining the learning material to the less proficient student might on itself contribute to enhanced understanding for both students (i.e., a mixed prior knowledge effect), and as a consequence, potential interaction effects of prior knowledge and teaching or learning methods might be confounded by such a mixed prior knowledge effect. Since we intended to examine potential interaction effects of prior knowledge and teaching or learning methods, we decided to split the prior knowledge groups and to leave the (also very interesting) mixed prior knowledge question for a future experiment.

In cognitive load theory, the development of propositional knowledge and conceptual understanding leads to enhanced knowledge structures about the subject matter in long-term memory (Van Merriënboer & Sweller, 2005). Students who have enhanced knowledge structures about the subject matter in their long-term memory are likely to experience lower cognitive load than their less knowledgeable peers when confronted with the subject matter or in an exam situation. Therefore, we expected that when learning statistics, low prior knowledge students experience more cognitive load than students having more prior knowledge. Further, we expected that low prior knowledge students who receive additional information – from the learning task or from a peer – experience less cognitive load during an exam than low prior knowledge students who do not receive additional information, whereas for students having more prior knowledge the trend goes in the other direction. Thus, the following three hypotheses were tested in this study:

- (1) Receiving additional information, from a peer or from a learning task, will enhance propositional knowledge and conceptual understanding only for low prior knowledge students.
- (2) Low prior knowledge students experience more cognitive load during task performance than students having more prior knowledge.
- (3) Receiving additional information, from a peer or from a learning task, leads to lower cognitive load during an exam only for low prior knowledge students.

5.2. Method

The current experiment compared the effects of self-explanation and explaining in pairs a couple of learning tasks on basic inferential statistics – for low prior knowledge students and for students having some more prior knowledge – on cognitive load, propositional knowledge, and conceptual understanding.

5.2.1. Participants and experimental design

A total of 84 bachelor students in psychology who had just failed their basic inferential statistics exam participated (grade 5 or lower on a ten point scale). Since the subject matter to be studied by the students (sampling distributions, the expected mean of a statistic, and hypothesis testing) formed an important part of the re-sit to be taken two months later, students were told that participation in the study could help them improve their results in the re-sit. Further, students were told that after having participated in the study, they would receive feedback about their performance in the study as well as advice on how to study for the re-sit. Thus, students participating had a stake in their performance in the current experiment.

Based on a median split of the exam scores, students were divided into two prior knowledge groups: the first group counted 42 students who had an almost sufficient grade to pass the exam (grade 4 or 5 on a ten point scale, while 6 is sufficient to pass) and a second group consisted of 42 students whose exam grade was lower (than 4 on a ten point scale). Within both groups, students were allocated at random to one of four possible treatment combinations. Having two experimental factors consisting of two levels each yields four treatment combinations. Experimental factors were whether or not the instruction in the learning task comprised additional information (i.e., instruction type) and whether students self-explained or explained in pairs of students (i.e., explanation type).

Thus, the experimental design was a 2 (instruction type) x 2 (explanation type) x 2 (prior knowledge) design, allowing us to test two- and three-way interactions. Six of the eight cells counted 10 students each (5 pairs), the other two cells counted 12 students each (6 pairs).

5.2.2. Materials

Materials were: (1) a text of four pages on basic inferential statistics, composed by the authors of the manuscript from chapters 4-6 of Moore et al. (2009), that had been subjected to a pilot-study for assessing its difficulty level and time required to read it properly, (2) one study assignment, consisting of two learning tasks, for each treatment combination, (3) a 45 minutes test consisting of two parts: one part measuring propositional knowledge and another part measuring conceptual understanding, and (4) a Dutch validated version of Paas' (1992) one-dimensional nine-point symmetrical category rating scale for cognitive effort.

5.2.3. Procedure

Participation in the study required 90 minutes from the students. The first 45 minutes formed the study session. The study session consisted of reading a study text (15 minutes) and performing two learning tasks related to that study text (30 minutes). In each of the learning tasks, students in the condition in which no additional information to the learning task was provided had to answer five open-ended questions to then integrate these answers into an argument proving a hypothesis to be true or false. One of these learning tasks can be found in Box 5.1.

Box 5.1.

One of the two learning tasks in the study assignment

To estimate an unknown population mean, we draw a random sample from that population. The sample mean is an estimator of the population mean.

Hypothesis: *The sample mean is not a random variable, since we draw only one sample and the mean in this sample has only one value.*

- [1] What is an estimator?
 - [2] Why is the sample mean an estimator?
 - [3] What is a random variable?
 - [4] What is a sampling distribution of an estimator?
 - [5] Why is the sampling distribution a probability distribution?
-

In the condition in which additional information to the learning task was provided, students did not have to answer these open-ended questions. Instead, they were given standard answers to these questions and they were instructed to integrate these answers into an argument. Further, students in the self-explanation condition performed the learning tasks individually, while students in the explanation to others (in pairs) condition performed the learning tasks in pairs. In the latter condition, the explanations occurred interactively. Since we chose to split up low prior knowledge students and students having more prior knowledge instead of pairing them off in the experimental conditions, we avoided situations in which one student explains and the other listens. In all the pairs in our experiment, both students gave multiple explanations to their partner. First, the theoretical constructs from the study text were discussed interactively, and when necessary, explained to each other. Consequently, the pairs typically formulated their argument for the learning task at hand together, and in the self-explanation condition they were inclined to provide joint answers to the open-ended questions to be integrated in the argument. Cognitive load was measured both after reading the study text and after performing the learning tasks by means of Paas' (1992) one-dimensional nine-point symmetrical category rating scale. This scale provides information with regard to the total cognitive load, which has been assumed to be a sum of the three types of cognitive load. To our knowledge, validated scales to measure the different types of cognitive load are still lacking (Beckmann, 2010).

After the 45 minutes study session (including two cognitive load measurements), students performed a 45 minutes test consisting of two parts. The first part consisted of a total of ten multiple-choice questions, typically embedded in a problem context, measuring conceptual understanding. These questions were derived from a pool of questions about the sampling distribution that had been used as exam questions in the previous years. An example of such questions can be found in Box 5.2. The second part of the test consisted of five of the ten open-ended questions used in the learning tasks (for examples see Box 5.1.). At the end of the test, cognitive load was measured a third time.

Box 5.2.

Example of a multiple choice item used to measure conceptual understanding

Three researchers draw each one random sample from the same population. Each of them wants to estimate the same unknown population mean of variable X . Suppose that this population mean is in fact equal to 21. In case A, a random sample of $N = 50$ is drawn and the sample mean is in fact equal to 21. In case B, a random sample of $N = 25$ is drawn and the sample mean is equal to 21. In case C, a random sample of $N = 10$ is drawn and the sample mean is equal to 24. In which of the three cases (A, B, C) is the expected value of the sample mean equal to the population mean?

[A] In case A

[B] In case B

[C] In case C

[D] In all cases (A, B, and C)

5.2.4. Data analysis

Two independent raters coded the open-ended questions by comparing students' short answers to Moore and McCabe's (2009) formulations. An incorrect answer was rated 0, a partly correct answer was rated 1, and a completely correct answer was rated 2. Initial agreement between the raters for the individual questions was high (Cohen's $\kappa = .87$), the correlation between the sum scores of the two raters was also high ($r = .95$), and the average difference in sum score very small. Examples of a completely correct, a partly correct, and an incorrect answer can be found in Box 5.3.

Box 5.3.

Example of how the open-ended questions were coded

Question: What is meant by the sampling distribution of an estimator?

Completely correct answer (i.e., 2 points): The sampling distribution of an estimator is the probability distribution of values of the estimator in all possible samples of the same size from the same population

Partly correct answer (i.e., 1 point): The sampling distribution of an estimator is the distribution of values taken by the estimator in more than one sample

Incorrect answer (i.e., 0 points): The sampling distribution of an estimator is the frequency distribution of (individual) scores that you find in your sample

For the multiple choice questions, incorrect choices were rated 0 and correct choices were rated 1. Given that the test consisted of five open-ended questions and ten multiple choice questions, both propositional knowledge and conceptual understanding were measured on a scale ranging from 0 to 10. We performed mixed linear regression having as fixed effects instruction type (answers to open-ended questions provided or not), explanation type (self or

joint), prior knowledge (low or higher), and their interactions, and a random slope for explanation type. Response variables were propositional knowledge, conceptual understanding, and cognitive load. We chose for this regression technique, since the experimental design was a multilevel design. When working in pairs on learning tasks, students' exam scores (within pairs) may be correlated more than students' exam scores from different pairs (between pairs). Although the 44 students in the self-explanation condition worked on the learning tasks individually and we therefore assumed 44 independent observations for each of the response variables, the 40 students in the explanation in pairs condition worked in pairs, meaning 20 and not 40 independent observations. A mixed linear model with a random slope for explanation type (i.e., self-explanation vs. explanation in pairs) provides a statistically correct solution for this problem.

5.3. Results

We had one hypothesis with regard to propositional knowledge and conceptual understanding and two hypotheses with regard to cognitive load, and we present the results separately.

5.3.1. Propositional knowledge and conceptual understanding

Our first hypothesis was that receiving additional information, from a peer or from a learning task, will enhance propositional knowledge and conceptual understanding only for low prior knowledge students. This hypothesis was partly confirmed for conceptual understanding, but not for propositional knowledge. Means and standard deviations (*SD*) of propositional knowledge score are presented in Table 5.1.

Table 5.1.
Means (and *SD*) and sample sizes (*n*) of propositional knowledge score

Group	Low prior knowledge	High prior knowledge
Propositional knowledge score (scale: 0-10)		
Self-explanation		
No additional info in task	2.83 (2.21) <i>n</i> = 12	4.00 (2.37) <i>n</i> = 12
Additional info in task	3.80 (1.99) <i>n</i> = 10	5.20 (1.48) <i>n</i> = 10
Explanation in pairs		
No additional info in task	3.80 (1.99) <i>n</i> = 10	4.60 (0.84) <i>n</i> = 10
Additional info in task	3.90 (1.85) <i>n</i> = 10	4.00 (1.89) <i>n</i> = 10

Table 5.2. presents betas and corresponding *F*-values and *p*-values with regard to the main effects of explanation type, instruction type, and prior knowledge group, as well as of the explanation type by prior knowledge group and instruction type by prior knowledge group interaction effect on propositional knowledge score. As the explanation type by instruction type interaction effect and the three-way interaction effect were not relevant for our hypothesis and these effects were not statistically significant, they are not presented in Table 5.2.

Table 5.2.

Unstandardized betas and corresponding F -values and p -values for the effects of explanation type, instruction type, and prior knowledge group on propositional knowledge score

Predictor	B	$F(1, 78)$	p -value
Intercept	3.02	38.160	< .001
Prior knowledge *	1.37	3.919	.051
Instruction **	0.55	0.869	.354
Explanation ***	0.55	0.869	.354
Explanation by prior knowledge	-0.81	0.942	.335
Instruction by prior knowledge	-0.21	0.065	.800

* 0 = low, 1 = high; ** 0 = no additional info, 1 = additional info; *** 0 = self-explanation, 1 = explanation in pairs

A model with only the main effect of prior knowledge group yielded a statistically significant effect: $F(1, 82) = 4.535$, $p < .05$. The unstandardized beta indicates that on average, low prior knowledge students develop less propositional knowledge than students having more prior knowledge. The findings with regard to propositional knowledge are not in line with our first hypothesis.

Means and standard deviations of conceptual understanding score are presented in Table 5.3.

Table 5.3.

Means (and SD) and sample sizes (n) of conceptual understanding score

Group	Low prior knowledge	High prior knowledge
Conceptual understanding score (scale: 0-10)		
Self-explanation		
No additional info in task	3.17 (1.59) $n = 12$	5.33 (1.83) $n = 12$
Additional info in task	4.60 (2.01) $n = 10$	4.40 (1.78) $n = 10$
Explanation in pairs		
No additional info in task	4.90 (1.91) $n = 10$	5.40 (1.43) $n = 10$
Additional info in task	5.50 (1.84) $n = 10$	4.30 (1.49) $n = 10$

Table 5.4. presents betas and corresponding F -values and p -values for the main effects of explanation type, instruction type, and prior knowledge group, as well as for the explanation type by prior knowledge group and instruction type by prior knowledge group interaction effect on conceptual understanding score. As the explanation type by instruction type interaction effect and the three-way interaction effect were not relevant for our hypothesis and these effects were not statistically significant, they are not presented in Table 5.4.

Although the explanation type by prior knowledge group interaction effect is not statistically significant, this is probably due to lack of statistical power. Tables 5.3. and 5.4. taken together, it

appears that only low prior knowledge students profit from additional information in the learning task and/or working in pairs (instead of alone) on learning tasks.

Table 5.4.

Unstandardized betas and corresponding *F*-values and *p*-values for the effects of explanation type, instruction type, and prior knowledge group on conceptual understanding score

Predictor	<i>B</i>	<i>F</i> (1, 78)	<i>p</i> -value
Intercept	3.30	59.714	< .001
Prior knowledge *	2.06	11.587	.001
Instruction **	1.13	4.177	.045
Explanation ***	1.33	5.306	.027
Explanation by prior knowledge	-1.34	2.709	.108
Instruction by prior knowledge	-2.05	7.352	.008

* 0 = low, 1 = high; ** 0 = no additional info, 1 = additional info; *** 0 = self-explanation, 1 = explanation in pairs

5.3.2. Cognitive load

We had two cognitive load hypotheses. Firstly, we expected that low prior knowledge students experience more cognitive load than students having more prior knowledge. Secondly, we hypothesized that receiving additional information (from a peer or from a learning task) leads to lower cognitive load during an exam only for low prior knowledge students. Only the first of these hypotheses was confirmed by the results. Means and standard deviations of cognitive load are presented in Table 5.5.

Table 5.6. presents betas and corresponding *F*-values and *p*-values for the main effects of explanation type, instruction type, and prior knowledge group, as well as for the explanation type by prior knowledge group and instruction type by prior knowledge group interaction effect on cognitive load. As the explanation type by instruction type interaction effect and the three-way interaction effect were not relevant for our hypothesis and these effects were not statistically significant, they are not presented in Table 5.6.

A model with only the main effect of prior knowledge group yielded a statistically significant effect: $F(1, 82) = 9.494, p < .01$. The unstandardized beta indicates that on average, low prior knowledge students experience more cognitive load than students having more prior knowledge.

Table 5.5.
Means (and *SD*) and sample sizes (*n*) of cognitive load

Group	Low prior knowledge	High prior knowledge
Cognitive load (scale: 1-9)		
After reading the text		
Self-explanation		
No additional info in task	5.42 (1.78) <i>n</i> = 12	4.00 (1.71) <i>n</i> = 12
Additional info in task	5.60 (2.01) <i>n</i> = 10	4.20 (1.32) <i>n</i> = 10
Explanation in pairs		
No additional info in task	4.80 (1.62) <i>n</i> = 10	4.10 (1.45) <i>n</i> = 10
Additional info in task	5.20 (1.90) <i>n</i> = 10	5.30 (1.34) <i>n</i> = 10
After reading the text		
Self-explanation		
No additional info in task	7.08 (1.78) <i>n</i> = 12	6.17 (1.90) <i>n</i> = 12
Additional info in task	7.30 (1.49) <i>n</i> = 10	7.00 (1.76) <i>n</i> = 10
Explanation in pairs		
No additional info in task	7.10 (0.99) <i>n</i> = 10	6.30 (1.70) <i>n</i> = 10
Additional info in task	7.10 (1.79) <i>n</i> = 10	5.60 (1.96) <i>n</i> = 10
After reading the text		
Self-explanation		
No additional info in task	6.83 (2.11) <i>n</i> = 12	6.33 (1.61) <i>n</i> = 12
Additional info in task	6.50 (1.66) <i>n</i> = 10	5.70 (2.36) <i>n</i> = 10
Explanation in pairs		
No additional info in task	6.70 (2.11) <i>n</i> = 10	5.90 (1.97) <i>n</i> = 10
Additional info in task	6.10 (1.66) <i>n</i> = 10	6.10 (1.37) <i>n</i> = 10

Table 5.6.
Unstandardized betas and corresponding *F*-values and *p*-values for the effects of explanation type, instruction type, and prior knowledge group on cognitive load

Predictor	<i>B</i>	<i>F</i> (1, 78)	<i>p</i> -value
Intercept	6.25	371.566	< .001
Prior knowledge *	-1.11	5.820	.018
Instruction **	0.14	0.132	.717
Explanation ***	-0.31	0.627	.431
Explanation by prior knowledge	0.34	0.371	.544
Instruction by prior knowledge	0.21	0.138	.711

* 0 = low, 1 = high; ** 0 = no additional info, 1 = additional info; *** 0 = self-explanation, 1 = explanation in pairs

5.4. Discussion

The results indicate a dominant prior knowledge group effect on both cognitive load and propositional knowledge, and differential effects of explanation type and instruction type for the

prior knowledge groups. In the remainder of this article, we discuss the findings as well as limitations of the current experiment and we suggest some implications for the teaching practice as well as well as for future research.

5.4.1. Limitations

The current study has a few limitations. To begin with, the current experiment used a convenience sample of students representing the lower range of the prior knowledge scale. Future studies could use a wider range of prior knowledge, for this may influence effect sizes of some of the effects examined in the current experiment. For example, the explanation type by prior knowledge group interaction effect, which is now a rather small effect, may be larger (and statistically significant) in studies that use a wider prior knowledge range.

A second limitation of the current experiment is that our conclusions with regard to cognitive load are based on Paas' (1992) one-dimensional nine-point symmetrical category rating scale. A pitfall of this technique is that it provides information only with regard to the total cognitive load, which has been assumed to be the sum of the three types of cognitive load. Therefore, some may criticize our interpretation by stating that the item we used to measure cognitive load does not allow us to draw straightforward conclusions with regard to potential instructional and prior knowledge effects on cognitive load. However, as Beckmann (2010) noted, valid techniques to measure the different types of cognitive load are lacking.

A third limitation of our experiment is that we did not videotape or audiotape the pair discussions or (think-aloud) self-explanations in the studying phase. We focused entirely on propositional knowledge, conceptual understanding, and cognitive load, and we have relatively little information on the actual implementation of the conditions by the students. Since the previously mentioned study by Leppink et al. (2011a; Chapter 2 of this thesis) demonstrates how think-aloud protocols can help us develop insight into thinking patterns during learning task performance in the domain of statistics, we recommend that future studies on self-explanation and explanation to others also focus on think-aloud processes.

5.4.2. Prior knowledge facilitates the development of propositional knowledge and moderates the development of conceptual understanding

The finding that low prior knowledge students experience more cognitive load and develop less propositional knowledge than students having more prior knowledge indicates that prior knowledge facilitates propositional knowledge development, independent of instruction type or explanation type. Thus, the findings do not suggest any expertise reversal effect for propositional knowledge.

With regard to conceptual understanding of statistics, the findings do suggest an expertise reversal effect. As expected, receiving additional information – by means of information in the learning task and/or by explaining in pairs – enhances conceptual understanding only among low prior knowledge students. Having low prior knowledge students explain the learning material in pairs and providing them with additional information in the learning assignment is optimal for helping them develop conceptual understanding of statistics. The fact that on average (3.17 on a 10-point scale) low prior knowledge students who self-explain the learning material and perform learning assignments without additional information hardly scored above what can be expected in the case of mere guessing (2.50 on a 10-point scale) illustrates that low prior knowledge

students who learn in these conditions hardly develop any conceptual understanding. However, when working in pairs on learning assignments and by providing them with additional information, low prior knowledge students can develop even more conceptual understanding than their peers who had more prior knowledge: the average of 5.50 in this group was (non-significantly) higher than among more proficient peers in any learning condition. For students having more prior knowledge, self-explanation or explaining in pairs hardly makes a difference with regard to conceptual understanding, and providing them with additional information in the learning assignment is likely to affect conceptual understanding negatively.

5.4.3. Implications for the teaching practice and for new research

Based on the findings with regard to propositional knowledge and conceptual understanding, we have a clear implication for the teaching practice and a few suggestions for future research.

Having students work in pairs on learning assignments to comprise additional information appears to be a good initiative for teachers to introduce the learning material and to elaborate on students' limited prior knowledge. However, once students have more prior knowledge, they should work on learning assignments that do not provide them with additional information, to avoid that such additional information actually hampers the development of their conceptual understanding. Thus, as students develop more knowledge of the learning material, they need less instructional guidance when performing learning assignments of a certain complexity level. For learning assignments of higher complexity level, students might need more guidance again until they have developed sufficient knowledge to learn from these more complex learning assignments. Given learning material of a certain complexity level, educational practice could strive to decrease guidance as the course proceeds. In the beginning, students could work in pairs on learning assignments that comprise additional information that elaborates on students' limited prior knowledge (e.g., the learning task presented in Box 5.1., but providing students with the answers to the open-ended questions). In a next phase, students could perform learning assignments individually, and self-explain the materials. Finally, students could engage in more unguided problem-solving. The latter situation is closer to real-life practice of statistics, in which statisticians, researchers, or other experts dealing with statistics, need to select relevant rules, propositions, or premises themselves, and apply their knowledge and understanding autonomously, without guidance.

Depending on the course aims, a combination of the self-guided and unguided problem-solving approach could be applied in the exam (e.g., two different types of exam problems). When applying the unguided approach in an exam following a course applying the peer-guided and self-guided approach, an interesting question is how students solve the problems in an exam, more specifically: will students still solve the problems by integrating relevant propositions? If yes, they demonstrate conceptual understanding, and that they can apply their knowledge and communicate with others in a way that makes sense.

Aleven and Koedinger (2002) demonstrated that the more unguided approach can be effective in geometry. Besides focusing on the (mixed) prior knowledge issue raised at the end of the previous paragraph, future studies could examine the effectiveness for the unguided approach for students varying in prior knowledge. The current experiment demonstrates that working in pairs on learning assignments that comprise additional information optimally supports low prior knowledge students in conceptual understanding development, whereas students

having more prior knowledge profit more from learning assignments that do not comprise additional information. Perhaps future studies will demonstrate that at a next level, students learn more from a more unguided approach.

Another question that could be answered by future studies is what the effects found in the current experiment would look like if each pair of students consisted of one low prior student and one student having more prior knowledge. In the current experiment, each pair consisted of either two low prior knowledge students or two students having more prior knowledge. It is possible that, when prior knowledge in pairs is mixed, the more proficient student explaining the learning material to the less proficient student leads to enhanced understanding for both students. The more proficient students then increase their understanding by explaining, thereby providing their less proficient peers with information they need to enhance their understanding. It would be very interesting to examine this question, since it can have major implications for teaching practice.

Chapter 6

Propositional manipulation in a statistics lecture

Submitted for publication

6.1. Introduction

Statistics is an important subject in many academic disciplines and curricula. A combination of factors contributes to the finding that many students develop only superficial understanding of this subject. To begin with, in many curricula the subject is given very limited time (Van Buuren, 2008), and part of that time is used to make students familiar with statistical software (Hulsizer & Woolf, 2009). Given that the domain of statistics is a complex knowledge domain that is characterized by abstract and hierarchical concepts that not always have a meaning outside the domain, avoiding the subject matter can easily lead to disorientation. Moreover, since for a proper understanding of the subject matter one has to understand a number of formulae and mathematical relationships, especially students with a non-mathematical background (e.g., many students in the social sciences) tend to avoid the subject matter (Broers, 2009). Appropriate teaching methods are needed to help students develop knowledge and understanding of statistics from the very start and step by step.

6.1.1. *Conceptual understanding of statistics*

A common way to introduce students into the study material is lecturing. When attending a lecture, students are confronted with important concepts and core ideas. Students have to first isolate these concepts and ideas by deriving their constituent elements to then structure these elements into schemata and gradually develop an integrated knowledge network (Novak, 2002). Knowledge of single statistical concepts and ideas is called propositional knowledge, whereas the development of an integrated network is referred to as conceptual understanding (Huberty et al., 1993). Although students usually develop propositional knowledge to a certain extent, many find it difficult to develop a conceptual understanding of the study material (Broers, 2009).

Propositional knowledge is a necessary but not sufficient condition for conceptual understanding (Leppink, 2011; Leppink et al., 2011a; Chapter 2 of this thesis). Developing conceptual understanding also involves self-explanation and argumentation (Aleven & Koedinger, 2002; Fischer, 2002; Knipfer et al., 2009). To avoid disorientation on the part of the students, they should be guided into these processes of self-explanation and argumentation (Broers & Imbos, 2005; Broers et al., 2005). Further, to avoid cognitive overload (Kalyuga & Hanham, 2010) and motivational constraints (Leppink, 2010), processes like self-explanation and argumentation should be shaped in such a way that students can directly elaborate on their prior knowledge (Kalyuga, 2009; Leppink et al., 2011a; Chapter 2 of this thesis). Learning statistics is a lengthy process requiring students' motivational states and development of knowledge and understanding of statistics to be taken into account. Research has shown that when one learns from an intrinsic motivation – a strong motivation from an internal desire to learn or perform – learning is more in-depth (Bruinsma, 2003), drop out is less likely, results are better (Ryan & Deci, 2000), curiosity is higher (Kuhl, 2000), one feels better in class (Levesque et al., 2004), and one is

more willing to cooperate and exchange information (Martens et al., 2004). These empirically supported assumptions form the core of the method of propositional manipulation (MPM), a teaching method for the statistics knowledge domain that has been developed by Broers (2002, 2008).

6.1.2. Propositional manipulation to help students build conceptual understanding

MPM consists of three steps. In the first step, the lecturer determines the study material and divides it into a limited number of statements referring to single statistical concepts and ideas. These statements are called propositions. The exact content and number of propositions depend on the content of the study material and should be aligned to students' prior knowledge of the study material. The lecturer formulates questions, each referring to one proposition. Examples of propositions and questions referring to these propositions are presented in Box 6.1.

Box 6.1.

Examples of propositions and questions referring to these propositions

Proposition 1: Robustness against violation of an assumption means that if the assumption is not met, the robust estimator or test will still have a reasonably small bias.

Question 1: What is robustness against violation of an assumption?

Proposition 2: One of the assumptions underlying one-way between-subjects ANOVA, the assumption of normality, is that the dependent variable follows a Normal distribution in each of the populations that we draw a sample from.

Question 2: What is, in the context of one-way between-subjects ANOVA, meant by the assumption of normality?

Proposition 3: One of the assumptions underlying one-way between-subjects ANOVA, the assumption of homogenous population variances, is that the dependent variable has an equal variance in the populations that we draw a sample from.

Question 3: What is, in the context of one-way between-subjects ANOVA, meant by the assumption of homogenous population variances?

Proposition 4: According to the Central Limit Theorem, if a population we draw a sample from does not follow a Normal distribution, the sampling distribution of our sample mean (or, more general: the sampling distribution of a sum or average) more and more approaches a Normal distribution as the size (n) of our sample increases.

Question 4: What is the Central Limit Theorem?

The core idea is that if the lecturer wants students to learn the propositions presented in Box 6.1., (s)he has to formulate a question for each of the propositions. By having the lecturer determine and decompose the study material this way, students have more cognitive resources available for learning.

In the second step of MPM, students are instructed to answer the questions formulated in the first step. Students are provided with the questions, not the actual propositions. The propositions

are taught to the students during the lecture and they can be found in the literature to be studied. Students are stimulated to self-explain the study material and they are guided into this process of self-explanation by means of the questions. The rationale behind this second step of MPM is that students develop the propositional knowledge they need to build conceptual understanding.

The aim of the third step in MPM is to help students build conceptual understanding. Students are instructed to perform a series of MPM learning tasks. In an MPM learning task, students have to relate and integrate a number of propositions into an argument that proves a given hypothesis to be either true or false. In contrast to propositions, the hypothesis typically comprises multiple statistical concepts and ideas. When instructing students to analyze and evaluate such a hypothesis, the lecturer can stimulate them to self-explain at a higher stage of complexity, namely by instructing them to create an argument that structures and relates propositions in such a way that students begin to understand why the hypothesis is true or false. Thus, MPM aims to help students develop conceptual understanding by guiding them into self-explanation at two different stages: first, at the stage of propositions (statements referring to single statistical concepts and ideas), and subsequently, at the stage of more complex problems that comprise a set of relevant propositions. It is the repeated formulation and practice of such arguments that should help students develop conceptual understanding of the study material.

The propositions have been chosen by the lecturer in the first step. For each proposition, the lecturer formulates one question (as presented in Box 6.1.). In the second step, students discover the propositions by answering the questions. In the third step, the lecturer provides the students with a hypothesis and attaches a limited number of propositions – formed as questions – to this hypothesis. Like for the propositions, the complexity level and formulation of the hypothesis depends on the learning goals of the course. Consider the example, one of the learning tasks used in the current study, presented in Box 6.2.

MPM stimulates students to engage in meaningful learning, as it stimulates them to self-explain the elements underlying the more complex hypothesis. Students are instructed to formulate an argument to prove the truth or falsity of the hypothesis based on the answers to the questions and their interrelationships. Students are not expected to learn the propositions through the formulation of arguments. The latter is supposed to help students build conceptual understanding, and the development of propositional knowledge is a necessary condition for conceptual understanding.

Propositional knowledge, self-explanation, and argumentation are prerequisites for the development of conceptual understanding. MPM encompasses these prerequisites. The lecturer determines the propositions and guides students into self-explanation of these propositions. At the next stage, the lecturer manipulates a relevant set of propositions in a learning task (see Box 6.2. for an example) and stimulates students to self-explain once again and to reason why the hypothesis (comprising the relevant set of propositions) is true or false.

Box 6.2.

One of the MPM learning tasks used in the current experiment

Contextual information: Suppose we run an experiment. A total of 90 participants is allocated randomly to one of three experimental treatment conditions:

- Experimental condition A (30 participants)
- Experimental condition B (30 participants)
- Control condition (30 participants)

Hypothesis: *Violation of the assumption of normality or the assumption of homogenous population variances does in our case not necessarily lead to largely biased test results.*

- [1] What is robustness against violation of an assumption?
 - [2] What is, in the context of one-way between-subjects ANOVA, meant by the assumption of normality?
 - [3] What is, in the context of one-way between-subjects ANOVA, meant by the assumption of homogenous population variances?
 - [4] What is the Central Limit Theorem?
-

6.1.3. The current study: propositional manipulation as a lecturing method

MPM is an instructional method for statistics education, based on the assumption that structuring the material to be learned and guiding students into self-explanation can (1) adjust cognitive load to their statistics proficiency level, (2) diminish avoidance behavior when confronted with the subject matter, (3) enhance their motivation to learn, and (4) stimulate them to engage in meaningful learning, so that conceptual understanding can be developed by integrating cognitive schemata into an organized knowledge network.

Although MPM was originally developed for the individual student (Broers, 2002) and all studies reported thus far focused on the potential of MPM for the individual student, suggestions have been made to apply MPM in interactive learning settings (Broers, 2008). The current study examined the potential of MPM in one such an interactive learning setting, namely in an interactive lecture. Previous studies reported that in individual learning settings MPM appears to be fruitful rather for students who have more prior knowledge (Leppink et al. 2011a, 2011b; Chapters 2 and 4 of this thesis). However, another recently conducted study suggests that studying and performing MPM learning tasks in pairs of students can compensate partly for prior knowledge deficiencies (Leppink et al., 2012; Chapter 5 of this thesis). If it turns out that MPM is equally fruitful for students from different prior knowledge levels when applied in an interactive lecture setting, a suggestion for the educational practice is that an interactive lecture in MPM format could help students raise their knowledge level such that performance of MPM learning tasks individually or in pairs of students will help them develop further their conceptual understanding of the study material.

Thus, our research question was whether students' prior knowledge affects the potential of MPM for individual students when applied in an interactive lecture. A potential advantage of an interactive lecture setting is that students who have little to no prior knowledge of the study material can benefit from the knowledge from their more knowledgeable peers. Assuming that

the latter knowledge is additional information that can contribute to the individual student's prior knowledge (Johnson et al., 2007), we hypothesized that low prior knowledge students and students having more prior knowledge can profit from MPM equally well when this method is applied in an interactive lecture. More specifically, we hypothesized that motivation to learn is elevated by MPM independent of prior knowledge, and that low prior knowledge students do not differ from their more knowledgeable peers in the extent to which they have developed conceptual understanding after having performed jointly a series of MPM learning tasks during an interactive lecture. Evidence confirming this hypothesis would indicate that MPM should be used in an interactive lecturing setting before using it in an individual setting, since in individual settings, at a relatively early stage, MPM appears to be potentially fruitful rather for high prior knowledge students (e.g., Leppink et al., 2011b; Chapter 4 of this thesis).

6.2. Method

The current study examined the potential effects of the method of propositional manipulation (MPM) as a lecturing method on motivation to learn and conceptual understanding of statistics.

6.2.1. Participants and experimental design

A total of 98 bachelor students in psychology who failed the exam of their second inferential statistics and were preparing for the re-sit registered to participate. To create a naturalistic setting, students were offered to attend a series of lectures covering the different topics in the course they had to re-sit. The lecture on the first topic, one-way and two-way between-subjects analysis of variance (ANOVA), was given in two different groups. Students were told that this was decided to increase the number of students that could participate in the lecture. To avoid that students who attended the first lecture would interact considerably with the students who were yet to attend the lecture or that students in one group had more days to revise any study material before the lecture, the lectures were given at the same day (11 A.M. till 1 P.M. and 1.30 P.M. till 3.30 PM respectively).

Students were allocated randomly to either of two groups. Students did not know they would participate in an experiment. Given that attending these lectures could increase students' chances at the re-sit a three weeks later, students participating had a stake in their performance in our experiment. The subsequent lectures were not part of the experiment, students attended those as one group.

Since we aimed to create a setting as naturalistic as possible, lecture attendance – as in the original course the students attended – was not mandatory. Thus, students were free to decide whether or not they would attend this lecture or only the next lecture(s). As a result, a total of twenty-seven registered students did not participate in our experiment (and most of them also did not attend any of the subsequent lectures in the end). Since students did not know they would participate in an experiment, this drop-out was not a consequence of any experimental treatment.

Topic, content, lecturer, and duration of both lectures were exactly the same, and in both lectures five true/false hypotheses were presented. Students in the first lecture ($n = 42$) discussed interactively the truth or falsity of each hypothesis. This group served as control group. In the second lecture (MPM group, $n = 29$), this interactive discussion was structured by presenting a number of short open-ended questions along with each hypothesis.

6.2.2. Materials

Materials were: (1) an SPSS (v19) output of a non-orthogonal two-way between-subjects ANOVA, (2) lecture slides for each of the treatment groups, (3) the intrinsic motivation inventory (IMI; Ryan, 1982; Ryan, Mims, & Koestner, 1983) to examine motivation to learn, and (4) a 60-minutes test consisting of twelve multiple choice questions with three alternatives, and various SPSS (v19) outputs of one-way and two-way between-subjects ANOVA pertaining to these multiple choice questions. The test items were embedded in a problem context and – like the hypotheses discussed during the lecture – comprised multiple statistical concepts and ideas. The items were taken from (re-sit) exams from previous years. An example of such an item is provided in Box 6.3.

Box 6.3.

Example of a multiple choice item in our experiment

Contextual information and SPSS (v19) output: left out here.

Question: two of the assumptions underlying one-way between-subjects ANOVA is that the dependent variable follows a Normal distribution in each of the populations we draw a sample from, and that the dependent variable has an equal variance in each of these populations. According to the findings presented:

[A] Both assumptions are violated significantly

[B] One of both assumptions is violated, but in our case the test is robust against this violation

[C] One of both assumptions is violated, and we would better not perform ANOVA here

The IMI is a multidimensional measurement device intended to assess participants' subjective experience related to a target activity in experiments. For a total of 45 statements, participants have to indicate on a 7-point scale (1 = not true at all, 7 = very true) how true the statement is for them. This device has been used in various experiments on intrinsic motivation and self-regulation (e.g., Deci, Eghrari, Patrick, & Leone, 1994; Ryan, 1982; Ryan, Koestner, & Deci, 1991), and has good reliability and validity (McAuley, Duncan, & Tammen, 1989; Markland & Hardy, 1997; Tsigilis & Theodosiou, 2003). Another reason why we chose for the IMI is because it comprises seven subscales, each referring to a different aspect of intrinsic motivation (e.g., perceived competence, autonomy or perceived choice, and relatedness). In our experiment, the target activity was the lecture. In other words, in the MPM group, students' IMI scores pertained to the structured discussion of each of the five true/false hypotheses (i.e., structured by the open-ended questions as in an MPM learning task), whereas in the control group students' IMI scores pertained to the unstructured discussion of these hypotheses (i.e., without the open-ended questions).

6.2.3. Procedure

Once students registered for participation, they were asked to fill in the IMI. Target activity was the course they just attended. Students' IMI and subscales' scores as well as students' exam performance (a score of maximum ten out of eighteen for students who failed the exam) served baseline measurements and in randomization checks.

The 60-minutes lecture was about one-way and two-way between-subjects ANOVA. Consecutively, the following five topics were treated: (1) assumptions underlying one-way and

two-way between-subjects ANOVA, (2) eta-squared as a measure of effect size, (3) the *F*-ratio, (4) multiple comparisons, and (5) non-orthogonal two-way designs and correction for confounding. Each of the topics covered ended with a true/false hypothesis. In both groups, students discussed interactively the truth or falsity of each of these statements. To avoid that many students responded simultaneously, the lecturer led the discussion. In the MPM group, the discussion around each of the hypotheses was structured around a number of open-ended questions. The open-ended questions and the hypothesis were presented simultaneously. Students were first instructed to answer each of the questions, and these answers could then be used to argue why the hypothesis displayed on the lecture slide was true or false. In the control group, each of the hypotheses was presented without any open-ended questions. The interactive discussion in this lecture group was guided by students' responses. In both groups, the lecture lasted exactly 60 minutes. At the end of the 60 minutes, students were asked to fill in the IMI again. This time, the target activity was the lecture they just attended.

During the 60 minutes following, the twelve items multiple choice test was administered. As each item comprised multiple statistical concepts and ideas and required students to select item-relevant information from extensive SPSS outputs, nearly all students needed between 50 and 60 minutes to complete the test. All students managed to complete the test within 60 minutes, without having to guess the final item(s).

6.2.4. Data analysis

Multivariate analysis of variance (MANOVA) and subsequent one-way ANOVAs were performed: (1) for randomization checks, treating students' IMI scores and IMI subscales' scores at baseline as correlated response variables and experimental treatment condition (i.e., MPM lecture vs. control group lecture) as independent factor, and (2) for the effect of MPM on motivation to learn, treating differences between students' IMI scores and IMI subscales' scores after the lecture and before the lecture as correlated dependent variables. Correlational analysis revealed that students' IMI subscales' scores at the time of registration for participation were more or less independent of students' exam performance, except for one subscale (effort / importance, $r = .22$, $p < .07$). The correlation between students' overall IMI score at the time of registration and students' exam performance was also close to zero ($r = .03$, $p > .80$). Therefore, students' exam performance was not included in the MANOVA for randomization checks. A separate one-way ANOVA was performed for the latter.

For the effect of MPM on conceptual understanding development (i.e., students' scores on the multiple choice test), analysis of covariance (ANCOVA) was performed, treating experimental treatment condition as independent factor. To examine potential prior knowledge effects and prior knowledge by experimental treatment condition interaction effects, students' initial exam scores were treated as covariate. Further, overall IMI scores and scores on the IMI subscales were considered as covariates to explore the potential effects of motivational variables on conceptual understanding development.

6.3. Results

We hypothesized that motivation to learn is elevated by MPM independent of prior knowledge, and that low prior knowledge students do not differ from their more knowledgeable peers in the extent to which they have developed conceptual understanding after having performed jointly a

series of MPM learning tasks during an interactive lecture. We first present the outcomes of a randomization check (paragraph 6.3.1.), and subsequently, we present the findings with regard to motivation to learn (paragraph 6.3.2.) and conceptual understanding (paragraph 6.3.3.).

6.3.1. Randomization checks

The groups were comparable with regard to both average exam performance and the motivational variables studied. The lecture groups did not differ significantly in average initial exam performance, $F(1, 69) = .365, p > .50$, and MANOVA revealed that the groups did not differ significantly in average motivation, $F(7, 63) = 0.915, p > .50$. Table 6.1. presents means and standard deviations (*SD*) on overall IMI score and scores on the IMI subscales at the time of registration per lecture group. Table 6.2. presents *F*-ratios and corresponding *p*-values of the subsequent one-way ANOVAs on lecture group differences in overall IMI score and scores on the IMI subscales.

Table 6.1.

Means (and *SD*) on overall IMI score and scores on IMI subscales at the time of registration per lecture group

Experimental treatment condition	Control lecture ($n = 42$)	MPM lecture ($n = 29$)
Overall IMI and IMI subscales' scores (scale: 1-7)		
Overall IMI	3.74 (0.56)	3.63 (0.69)
Interest / enjoyment	2.99 (0.95)	3.08 (1.04)
Perceived competence	2.57 (0.97)	2.59 (0.86)
Effort / importance	4.42 (1.27)	4.23 (1.62)
Pressure / tension	4.26 (1.33)	4.06 (1.58)
Perceived choice	2.50 (1.23)	2.62 (1.26)
Value / usefulness	5.26 (1.15)	4.85 (1.25)
Relatedness	4.18 (1.14)	4.01 (0.87)

Table 6.2.

F-ratios and corresponding *p*-values of the subsequent one-way ANOVAs on lecture group differences in overall IMI score and scores on the IMI subscales at the time of registration per lecture group

Response variable	Cohen's <i>d</i>	$F(1, 69)$	<i>p</i> -value
Overall IMI and IMI subscales' scores (scale: 1-7)			
Overall IMI	0.17	0.489	.49
Interest / enjoyment	0.09	0.129	.72
Perceived competence	0.03	0.012	.91
Effort / importance	0.13	0.304	.58
Pressure / tension	0.14	0.333	.57
Perceived choice	0.10	0.172	.68
Value / usefulness	0.34	2.008	.16
Relatedness	0.16	0.445	.51

6.3.2. MPM does not improve motivation to learn

Table 6.3. presents means and standard deviations on overall IMI score and scores on the IMI subscales after the lecture per lecture group.

Table 6.3.
Means (and *SD*) on overall IMI score and scores on IMI subscales after the lecture per lecture group

Experimental treatment condition	Control lecture (<i>n</i> = 42)	MPM lecture (<i>n</i> = 29)
Overall IMI and IMI subscales' scores (scale: 1-7)		
Overall IMI	4.90 (0.56)	4.96 (0.50)
Interest / enjoyment	3.89 (0.95)	4.22 (0.92)
Perceived competence	3.26 (1.15)	3.20 (1.25)
Effort / importance	5.34 (0.94)	5.08 (1.04)
Pressure / tension	5.26 (1.32)	5.26 (1.07)
Perceived choice	5.24 (1.17)	5.06 (1.02)
Value / usefulness	6.14 (1.06)	6.57 (0.44)
Relatedness	5.16 (0.67)	5.30 (0.65)

MANOVA revealed a non-significant overall effect of experimental treatment condition (i.e., lecture group) on motivation, $F(7, 63) = 1.695$, $p > .10$. Table 6.4. presents *F*-ratios and corresponding *p*-values of the subsequent one-way ANOVAs on lecture group differences in IMI difference and differences in scores on the IMI subscales.

Table 6.4.
F-ratios and corresponding *p*-values of the subsequent one-way ANOVAs on lecture group differences in IMI difference and differences in scores on the IMI subscales

Response variable	Cohen's <i>d</i>	<i>F</i> (1, 69)	<i>p</i> -value
Overall IMI and IMI subscales' scores (scale: 1-7)			
Overall IMI	0.11	0.940	.34
Interest / enjoyment	0.28	0.591	.45
Perceived competence	0.05	0.078	.78
Effort / importance	0.26	0.042	.84
Pressure / tension	< 0.01	0.320	.57
Perceived choice	0.16	0.601	.44
Value / usefulness	0.49	7.127	< .01 *
Relatedness	0.21	1.198	.28

* the exact *p*-value was .009 and after correction for multiple testing no longer statistically significant at the .05 level

Although in the subsequent one-way ANOVAs, the group difference in value / usefulness was statistically significant, $F(1, 69) = 7.127$, $p < .01$, after Bonferroni correction for multiple testing

this difference is no longer statistically significant. Besides, due to restriction of range in the value / usefulness scores this effect is difficult to interpret. In both lecture groups, the majority of students evaluated the lecture as very useful (i.e., the majority of students scored the maximum score of 7 or close to 7 on this variable).

6.3.3. MPM can enhance conceptual understanding development in statistics

One-way ANOVA treating experimental treatment condition as independent factor revealed a significant treatment effect, $F(1, 69) = 5.753, p < .05$ (Cohen's $d = 0.56$): on average, students in the MPM lecture group ($M = 7.66, SD = 1.70$) performed better than students in the control group ($M = 6.76, SD = 1.43$). This indicates a medium size effect of MPM.

To examine potential prior knowledge effects and prior knowledge by experimental treatment condition interaction effects, students' initial exam scores were treated as covariate. Since previous studies reported an interaction between instructional method and students' prior knowledge levels, we first ran a model including experimental treatment condition, students' initial exam scores, and the interaction effect of these two. This yielded a non-significant interaction effect, $F(1, 67) = 2.399, p > .45$. A subsequent ANCOVA model including experimental treatment condition and students' initial exam scores indicates that the effect of initial exam score was not statistically significant, $F(1, 68) = .966, p > .30$.

Correlational analysis revealed that students' IMI subscales' scores both at the time of registration were more or less independent of students' conceptual understanding score, except for one subscale (relatedness, $r = .21, p < .08$). The correlation between students' overall IMI score at the time of registration and students' conceptual understanding score was also low ($r = .13, p > .25$), and a similar correlation was found between students' overall IMI score after the lecture and students' conceptual understanding score ($r = .09, p > .45$). After the lecture, the correlation between perceived competence and conceptual understanding score was statistically significant ($r = .27, p < .05$), the correlation between value / usefulness and conceptual understanding score was low and not statistically significant ($r = .13, p > .25$), and the other five IMI subscales had correlations with conceptual understanding close to zero.

Thus, MPM and higher perceived competence contributed to higher conceptual understanding score, while MPM did not lead to significantly increased perceived competence. In other words, perceived competence does not appear to mediate the effect of MPM on conceptual understanding score. To examine further potential joint effects of MPM and perceived competence, we ran a model including experimental treatment condition, students' perceived competence scores after the lecture, and the interaction effect of these two. This yielded a non-significant interaction effect, $F(1, 67) = 1.782, p > .15$. A subsequent ANCOVA model including experimental treatment condition and students' perceived competence scores after the lecture indicates a positive and statistically significant effect of perceived competence on conceptual understanding score, $F(1, 68) = 5.969, p < .05$, as well as a significant effect of MPM on conceptual understanding score, $F(1, 68) = 6.485, p < .05$ (Cohen's $d = 0.61$). The latter indicates a medium size effect of MPM on conceptual understanding score.

Taken together, the findings suggest that MPM does not improve motivation to learn but can enhance conceptual understanding development of students.

6.4. Discussion

Our research question was whether students' prior knowledge affects the potential of MPM for individual students when applied in an interactive lecture. A potential advantage of an interactive lecture setting is that students who have little to no prior knowledge of the study material can benefit from the knowledge from their more knowledgeable peers. Our hypothesis that motivation to learn is elevated by MPM cannot be confirmed, but we did find that low prior knowledge students and students having more prior knowledge can profit from MPM equally well when this method is applied in an interactive lecture setting.

Previous studies (e.g., Leppink et al. 2011b; Chapter 4 of this thesis) reported that in individual learning settings MPM appears to be fruitful rather for students who have more prior knowledge, while from our current experiment it appears equally fruitful for students from different prior knowledge levels when applied in an interactive lecture setting. This raises the question whether MPM could be applied in a statistics course first in an interactive lecture to introduce the subject matter, and consecutively in smaller group or individual learning settings. The lecture could help students raise their knowledge level such that performance of MPM learning tasks individually will help them develop further their conceptual understanding of the study material. Future research could test this option, for example in a two-way factorial design giving students a lecture in MPM format or not and having students study individually in MPM format (i.e., performing a series of MPM learning tasks) or study in a more unguided way. Students' prior knowledge, and possibly its interactions with treatment conditions as well, could then be included in the model as covariates.

Another way to study the effect of MPM on students' conceptual understanding development when applied in a lecture setting is to provide students with materials that they are instructed to complete during the lecture. In the MPM group, students then perform a series of MPM learning tasks during the lecture, whereas students in the control group perform learning tasks about exactly the same content (and about exactly the same lecture) but in a different format. For example, in a lecture about ANCOVA, students in the MPM group receive a true/false hypothesis together with a number of open-ended questions and the instruction to answer the open-ended questions (during the lecture) and then formulate an argument that proves the presented hypothesis to be true or false, whereas students in the control group receive only the statement with the instruction to reason in their own words why the statement is true or false.

A limitation of our experiment may be that the students in our study were not the most proficient students. The range of prior knowledge was quite limited. On the other hand, the prior knowledge range in our experiment may be more in line the prior knowledge range among students at the beginning of a course than a prior knowledge range including students who already passed the exam. Therefore, we do not consider our prior knowledge range as a serious shortcoming of our study. However, it can be questioned whether this argument holds for motivation to learn as well. Some may argue that, by sampling from the students who failed the exam, the students in our experiment form a sample from the lower range of motivation to learn. On the other hand, given that attending these lectures could increase students' chances at the re-sit a three weeks later, students participating had a stake in their performance in our experiment.

Some may argue that a volunteer bias limits the implications of our findings. Although we are aware of the possibility of this bias, we created a learning setting that is comparable to a learning

setting in a regular course. Students who were preparing for an exam (here: re-sit) attended a series of lectures, of which the first lecture was given to two different groups of students for an experimental purpose that none of the students were aware of. Students did not have any financial benefit from participating in our study, the only advantage for them was a better preparation for the re-sit than a preparation without lectures. This is similar to a regular course setting: students who do not attend any lectures are more likely to fail the exam at the end of the course than students who do attend the lectures.

Chapter 7

Guided problem-based learning of statistics

Submitted for publication

7.1. Introduction

Problem-based learning (PBL; Barrows, 1984; 1996) is applied in a wide range of academic disciplines across the world. Six general principles of PBL (Schmidt, Van der Molen, Te Winkel, & Wijnen, 2009) are: (1) a problem is the starting point for new learning, (2) PBL supports collaboration through a small group problem-based discussion, (3) the group is guided by a tutor, whose role is to facilitate the group discussion by stimulating knowledge elaboration, (4) PBL programs comprise fewer lectures than most other curricula, (5) based on a group discussion that helps students activate their prior knowledge, students are encouraged to actively search for literature related to the problem, and (6) students should receive sufficient time for self-study.

7.1.1. Instructional guidance in problem-based learning

One of the core elements of PBL is knowledge elaboration (Springer, Stanne, & Donovan, 1999). Knowledge elaboration can be stimulated by relating the group discussion to students' prior knowledge (Lee, Lin, Tsou, Shiau, & Lin, 2009). However, it is to be questioned to what extent PBL is effective for students who barely have prior knowledge of the subject or even have incorrect prior 'knowledge' as is frequently the case in complex knowledge domains like statistics (Kaplan, 2006).

According to Kirschner, Sweller, and Clark (2006), PBL is an inappropriate format for this group of students. All problem-based searching relies heavily on students' working memory, and working memory capacity that is being used for searching cannot be used for learning. This is especially the case for students who have little to no prior knowledge of the domain (Willoughby, Waller, Wood, & MacKinnon, 1993; Woloshyn, Pressley, & Schneider, 1992). From this perspective, it is not surprising that Bude, Imbos, Van de Wiel, and Berger (2009) found that especially in complex knowledge domains like statistics a more guiding or directive tutor is more likely to yield optimal learning results in PBL groups than a tutor who remains on the background and only asks questions that are supposed to stimulate prior knowledge activation.

Another way to provide more instructional guidance in PBL group discussions, that to our knowledge has not yet been examined, is to provide students with a number of specific learning goals on the subject. In the words of Kirschner and colleagues (p. 77): "The goal of instruction is rarely simply to search for or discover information. The goal is to give learners specific guidance about how to cognitively manipulate information in ways that are consistent with a learning goal, and store the result in long-term memory." The need for specific guidance may be especially necessary when students are introduced into complex knowledge domains. For example, the statistics knowledge domain consists of many hierarchically structured concepts. Since beginning students usually lack prior knowledge of the domain, they are unlikely to formulate appropriate and specific learning goals that help them in their subsequent search for information in the myriad of concepts and ideas in the literature and that together cover the subject to be studied.

There appears to be a need for several specific learning goals – formulated by the instructor and each referring to different concepts and ideas – rather than a limited number of broader learning goals formulated by the students themselves. Further, in complex knowledge domains that are characterized by hierarchically structured concepts, learning goals at different hierarchical levels may be required.

7.1.2. The current experiment: problem-based learning and guided problem-based learning

Statistics is a complex knowledge domain that is considered to be an indispensable part of many academic curricula, including psychology, health sciences, and medicine. At some point or other, students in these curricula are likely to encounter one or more mandatory courses in statistics. Given that these students typically have limited (and partly incorrect) prior knowledge of statistical methods, and frequently lack the ability to understand the interrelationships between statistical concepts and ideas (i.e., conceptual understanding, Broers, 2009), the need for specific and hierarchical learning goals formulated by the instructor appears to be obvious for this domain. Not only may specific and hierarchical learning goals help students develop more conceptual understanding of statistics, it may also help them become aware of the value and usefulness of this subject for their academic development. Students who lack conceptual understanding of statistics frequently approach the subject with dislike and apprehension (Gal & Ginsburg, 1994) and complain that they do not understand why studying this difficult and boring subject is important for them.

From the theory and the literature we can expect that the required instructional guidance depends on students' prior knowledge as well as on the complexity of the knowledge domain. This may explain why PBL is effective for many medicinal and social science related subjects but may be less effective in the domain of mathematics and statistics. Therefore, the aim of the current experiment was to compare – in terms of conceptual understanding of statistics as well as in terms of perceived value and usefulness of learning statistics – the classical PBL format, in which learning goals are the product of a problem-based discussion by a small group of students, and a guided PBL (GPBL) format in which the learning goals are formulated by the instructor to guide and structure group discussion.

Thus, two related research questions were examined: (1) does GPBL enhance students' conceptual understanding of statistics more than PBL? and (2) does GPBL enhance students' perceived value and usefulness of learning statistics more than PBL? Based on the theoretical considerations presented, we hypothesized that GPBL would enhance both students' conceptual understanding of statistics and students' perceived value and usefulness of learning statistics. Further, without specific hypotheses in mind, we explored students' motivation to learn statistics more generally (perceived value and usefulness is just one specific aspect), students' participation in the PBL and GPBL groups and whether this influenced conceptual understanding of statistics, and we explored students' as well as tutors' experiences concerning the PBL and GPBL group discussions.

7.2. Method

The hypotheses that GPBL will enhance students' conceptual understanding of statistics and also enhance students' perceived value and usefulness of learning statistics were tested in an experimental setup.

7.2.1. Participants and experimental design

University freshmen were recruited from the faculties of psychology and health sciences. These students were about to encounter their first inferential statistics course. They were informed that the subject matter treated in the study formed the core of that course, and that an individual feedback session would follow one week after their participation in the study. This created a naturalistic setting. Although the (G)PBL groups were planned at times when no regular educational activities were planned for these students, students who were willing to participate were asked to confirm their availability in the period in which the (G)PBL groups were planned. A total of 110 students confirmed their availability. These 110 students were allocated randomly to one of ten groups, and subsequently, the ten groups were allocated randomly to either PBL or GPBL. Each group originally counted eleven students. Sixteen students had to cancel their participation due to unexpected changes in their educational timetable. As students did not know before the studying session in what condition they would participate and they were not aware of any experimental treatment conditions, this drop-out was not a consequence of any experimental treatment.

Each treatment condition had five groups. Group size varied from eight to eleven students. Students did not know that there were different treatment conditions. Each of the groups was led by one of three tutors. Each tutor was a Research Master's student in psychology who mastered the subject very well and worked according to a tutor protocol that was designed for the study. Two tutors led four of the ten groups (two groups in each format), and the third tutor led the remaining two groups (one group in each format). An overview of the experimental design is presented in Table 7.1.

Table 7.1.
Schematic overview of the experimental design

Group code	Experimental treatment	Number of students	Tutor
G1	PBL	10	C
G2	PBL	9	B
G3	GPBL	9	A
G4	GPBL	10	C
G5	GPBL	9	B
G6	PBL	8	A
G7	PBL	10	C
G8	GPBL	8	A
G9	GPBL	10	C
G10	PBL	11	A

N.B.: tutors A and C each led two groups in each experimental treatment condition, while tutor B led only one group in each experimental treatment condition

Finally, although tutors were of course aware that there were two different instructional formats, they were blind about theoretical background and our expectations with regard to experimental treatment.

7.2.2. Materials

Materials were: (1) a ten multiple choice questions pretest on statistical reasoning, (2) a text of four pages on basic inferential statistics, composed by the authors of the manuscript that was used in previous studies as well, (3) one study task per treatment condition, (4) the intrinsic motivation inventory (IMI, Ryan, 1982; Ryan et al., 1983) to examine motivation to learn, and more specifically the subscale of perceived value and usefulness of the learning activity, and (5) a thirty minutes test on conceptual understanding of statistics, consisting of ten multiple choice questions with four alternatives of which one was correct. The test items were embedded in a problem context and comprised multiple statistical concepts and ideas. The items were taken from (or in some cases: based on) exams from previous years. An example of such an item is provided in Box 7.1.

Box 7.1.

Example of a multiple choice item in the current experiment

Contextual information was the same for all ten items and is left out here.

Another reason why Mickey has doubts about the size (N) her sample should have, is because she does not know whether the student population follows a Normal distribution on the motivational variable. In which of the following cases can Mickey be most secure that the distribution of possible values for the sample mean resembles a Normal distribution?

[A] If $N = 10$

[B] If $N = 25$

[C] If $N = 40$

[D] None of the previous alternatives is correct, because the form of this distribution cannot resemble a Normal distribution if Mickey's student population does not follow a Normal distribution

The IMI is a multidimensional measurement device intended to assess participants' subjective experience related to a target activity in experiments. For a total of 45 statements, participants have to indicate on a 7-point scale (1 = not true at all, 7 = very true) how true the statement is for them. This device has been used in various experiments on intrinsic motivation and self-regulation (e.g., Deci et al., 1994; Ryan, 1982; Ryan et al., 1991), and has good reliability and validity (McAuley et al., 1989; Markland & Hardy, 1997; Tsigilis & Theodosiou, 2003). Another reason why we chose for the IMI is because it comprises seven subscales, each referring to a different aspect of intrinsic motivation, and one of these scales is the perceived value and usefulness scale. In our experiment, students' scores on the IMI and subscales pertained to the activities in the PBL or GPBL group they participated in (i.e., group discussion prior to self-study, self-study, and group discussion after self-study).

7.2.3. Procedure

Once students registered for participation, they were asked to fill in the ten-item pretest on statistical reasoning. Each items consisted of four alternatives of which one was correct and one other alternative represented a typical misconception students typically have about the topic.

Students' sum score of correctly chosen alternatives served as randomization checks and as a covariate in our statistical analyses.

Subsequently, students participated in the tutorial group they were allocated to. The tutor leading the group did not participate actively in the group (s)he only guarded the time and asked a standard question as provided in the tutor protocol in moments the group discussion appeared to go into a dead end or the group discussion tended to shift into unrelated topics or into serious misconceptions about the topic. In none of the groups, the tutor provided answers or explanations to questions, (s)he was instructed by means of the questions in the tutor protocol to stimulate students (in a standardized way) to elaborate their knowledge. The questions formulated in the tutor protocol served to standardize the group process, so that confounding due to tutor or unequal treatment of groups could be ruled out in the data analysis. Some examples of questions in the tutor protocol are presented in Box 7.2.

Box 7.2.

Examples of questions in the tutor protocol

Examples of questions for the discussion prior to self-study

- [1] We can be sure that (some of) you know what is an arithmetic mean. Can someone explain how we calculate such a mean on the basis of a set of scores?
- [2] What does it mean that we draw at random ten students from the population?

Examples of questions for the discussion after self-study

- [1] Can someone explain what is a *p*-value?
 - [2] Can someone explain what a sampling distribution is?
 - [3] Can someone explain what a significance level is?
-

For all the groups, the general timeline was as follows: (1) thirty minutes discussion of a problem (which was the same for all groups), (2) thirty minutes self-study on the basis of the text composed by the authors, (3) thirty minutes discussion of the problem on the basis of the self-study, and (4) the thirty minutes test consisting of ten multiple choice questions with four alternatives of which one was correct. Group discussion before and after self-study were audio-taped and later transcribed verbatim for qualitative analysis.

In the PBL groups, students formulated learning goals themselves, based on the discussion of the problem before self-study. These learning goals then guided both self-study and the group discussion after the self-study period. In the GPBL groups, learning goals were presented along with the problem text, to help students activate their prior knowledge and to structure self-study as well as group discussion before and after self-study. That is, in the GPBL groups, students were confronted with the two learning tasks that are presented in Box 7.3.

To represent the hierarchical nature of the statistics knowledge domain, each of the two learning tasks comprised learning goals at the level of open-ended questions that referred to single statistical concepts and ideas and could be answered by means of short propositional statements. These propositional statements (i.e., propositional knowledge; Broers, 2002; Huberty et al., 1993) could then be integrated into an argument that proves the hypothesis – a statement

comprising multiple statistical concepts and ideas and links between these concepts and ideas – to be true or false. The latter step requires conceptual understanding, which is the ability to understand how statistical concepts and ideas are interrelated (Broers, 2009; Huberty et al., 1993). This method of first having students answer open-ended questions that can then help them to evaluate more complex hypotheses is also called the method of propositional manipulation (MPM; Broers, 2002, 2009). Although this method was originally developed for and examined within the context of individual learning (Broers, 2002; Leppink et al., 2011a; Chapter 2 of this thesis), a recent experiment suggests that, when learning in pairs, this method can enhance conceptual understanding of statistics even for students who have relatively little prior knowledge (Leppink et al., 2012; Chapter 5 of this thesis).

Box 7.3.

The learning tasks for the directive learning condition in the current experiment

Statement in learning task 1: *If the p -value equals 0.03 and the significance level is 0.05, the null hypothesis is false.*

Underlying (propositional) questions:

- [1] What is meant by null hypothesis and alternative hypothesis?
- [2] What is meant by expected mean of an estimator?
- [3] What can we say about the sampling distribution of the sample mean under the null hypothesis?
- [4] What exactly is a p -value?
- [5] What is the meaning of a significance level?

Statement in learning task 2: *In a large sample, the sample mean is a more precise estimator of the population mean than it is in a small sample.*

Underlying (propositional) questions:

- [1] What is the difference between sample mean and population mean?
- [2] Why is the sample mean an estimator of the population mean?
- [3] What is a sampling distribution of an estimator?
- [4] What is the relationship between the size of a sample and the variation in possible values for the sample mean?
- [5] What does statistical theory tell us about the form of the sampling distribution of the sample mean?

N.B.: the open-ended questions were the learning goals in the GPBL condition. The statements were discussed on the basis of how the students answered these open-ended questions for themselves during self-study and in the group discussion after self-study

Predefined learning goals (the learning tasks presented in Box 7.3. for GPBL) versus no predefined learning goals (PBL) was the only difference between the two treatments. In the group discussion prior to self-study, the learning goals in the GPBL groups were intended to activate relevant prior knowledge. In the group discussion after self-study, the statements were evaluated on the basis of how the students answered these open-ended questions for themselves during self-study and in the group discussion after self-study.

Completing the multiple choice test formed the last thirty minutes of the two-hour group meeting. One week after participation in the group, students returned to the laboratory, to

complete the same multiple choice test once again. Students had been told that the meeting one week later did not require any preparation from them, that they would only receive feedback on their performance at the end of the group meeting. Thus, students did not know that before returning to the laboratory that they would have to complete the test again. Students' second test scores enabled us to explore delayed recall and how much students would forget within the week after the meeting. It is of course possible that different treatment conditions have a different impact on students' forgetting. Given that in real academic curricula there is frequently about a week interval between consecutive PBL groups, and that in a complex knowledge domain like statistics the subject matter in the next week builds hierarchically on the subject matter of previous weeks, it is interesting to examine to what extent students still remember or understand certain concepts and ideas approximately one week after a group meeting.

7.2.4. Data analysis

To test our two hypotheses – that GPLB would enhance both students' conceptual understanding of statistics and students' perceived value and usefulness of learning statistics – comparisons were made at the level of groups with eight to eleven students. The 94 students participating in our study cannot be considered to be 94 independent units. Furthermore, the groups were divided over three tutors. To take the correlational nature of our units within groups into account, we tested our hypothesis by means of two linear multilevel regression models treating tutor and (G)PBL group as levels (this was the same for the multilevel regression models described later on), having as one dependent variable "students' conceptual understanding score" (i.e., Model 1) and one dependent variable students' "perceived value and usefulness of statistics" (i.e., Model 2). An overview of these and other multilevel regression models for data analysis is to be found in Table 7.2.

As the correlation between these two dependent variables was – for both treatment conditions – not consistent across groups (varying from $r = -.56, p = .094$ in one group to $r = -.08, p > .80$ and even $r = .46, p = .215$ in other groups) and not statistically significant in any of the groups, we did not include them as correlated dependent variables in one single model. The experimental treatment condition, students' prior knowledge, and their interaction between experimental treatment condition and students' prior knowledge were predictors/covariates in our model. For both treatment conditions, the skewness of the dependent variables was within the range of ± 0.5 , and the kurtosis was within the range of ± 1.0 , implying no serious departures from normality. The skewness and kurtosis of the prior knowledge variable were both within the range of ± 0.5 . Further, as no outliers or extreme values in any of these variables were detected, and standard deviations of scores on the dependent variables were comparable for the experimental treatment conditions (i.e., less than a factor 1.5 difference) indicating no serious departures from homoscedasticity, transformation of our dependent variables for further analysis was not needed.

Since we expected two main effects for experimental treatment condition (i.e., on conceptual understanding score in Model 1, and on perceived value and usefulness score in Model 2), we expected experimental treatment condition to be statistically significant but not the interaction between experimental treatment condition and students' prior knowledge.

Further, entering students' prior knowledge scores as such in a regression model, the regression coefficient for the experimental treatment represents treatment group differences at

zero prior knowledge instead of at average prior knowledge. Since we were interested in a potential main effect of experimental treatment condition, we needed to examine treatment group differences at average prior knowledge. To achieve this, we used centered prior knowledge score (instead of students' original prior knowledge scores) in our analyses (i.e., the average score was subtracted from each individual score, and as a result the average of the centered prior knowledge scores equals zero).

Table 7.2.
Overview of the five multilevel regression models for data analysis

Model	Purpose
1	To test our hypothesis of a main effect of GPBL over PBL on “students’ conceptual understanding score” and no significant interaction between experimental treatment condition (GPBL vs. PBL) and students’ prior knowledge (and this hypothesis was rejected, for the interaction effect was statistically significant).
2	To test our hypothesis of a main effect of GPBL over PBL and “students’ perceived value and usefulness of learning statistics” and no significant interaction (and this hypothesis was not rejected).
3	To explore whether experimental treatment condition (GPBL vs. PBL), students’ prior knowledge, and the interaction between experimental treatment condition and students’ prior knowledge could explain whether students would participate actively in a group discussion or not.
4	To explore the effect of active participation (i.e., active or not) on “students’ conceptual understanding score”.
5	A model comprising all predictors/covariates of Model 1 and Model 4 together.

We examined exploratively the effects of the abovementioned covariates on students’ scores on the IMI subscales other than the perceived value and usefulness scale, and we applied Bonferroni correction in our tests.

To explore (without hypothesis in mind) the learning goals (in the PBL groups), reasoning, and argumentation along the group discussions (in both PBL and GPBL groups), the transcribed protocols of the group discussions were subjected to content analysis. Further, at the end of the study, we asked each of the tutors to report independently their general impressions of both treatment conditions, and to evaluate differences between the treatment conditions with regard to (1) information-richness of the group discussion and the number of students attending the group discussion, (2) students’ motivation to learn, and (3) the development of knowledge and understanding by the students. This yielded independent insights into the group processes in PBL and GPBL.

Furthermore, in every group it was determined who contributed in the discussion – by providing explanations or answers and/or by asking questions – or not. Although we did not have any specific hypothesis about whether PBL and GPBL would differ in the proportion of students involved in the discussion, we intended to explore the effects of experimental treatment condition, students’ prior knowledge, and their interaction effect on the likelihood of

participating in the discussion. We performed logistic multilevel regression analysis with being active or not as dependent variable and using experimental treatment condition, students' prior knowledge, and their interaction effect as covariates in the model (i.e., Model 3). Again, for the interpretation of a potential main effect of experimental treatment condition, centered prior knowledge scores were used. Subsequently, we performed linear multilevel regression analysis with students' conceptual understanding of statistics as dependent variable and being active or not as predictor (i.e., Model 4).

Finally, we performed a linear multilevel regression model for students' conceptual understanding as dependent variable with covariates: experimental treatment, centered prior knowledge, the interaction between experimental treatment condition and centered prior knowledge, and being active or not (i.e., Model 5). This linear regression model provides information with regard to the direct effects of the experimental treatment, students' prior knowledge, and the interaction between these two, while the linear regression model without the covariate of being active or not provides information with regard to the total effects (i.e., direct and indirect) of these covariates. Since being active or not in a group discussion is potentially causally influenced by experimental treatment, being active or not is a mediator and part of the experimental treatment effect may be mediated by being active or not.

7.3. Results

We consecutively present: (1) the results with regard to students' prior knowledge and randomization checks, (2) the findings with regard to students' perceived value and usefulness, and more generally, their motivation to learn, (3) the findings on students' conceptual understanding scores, and (4) a report on students' learning goals in the PBL groups, and students' argumentation in the PBL and GPBL groups.

7.3.1. Randomization checks

The treatment conditions did not differ significantly in average prior knowledge, $F(1, 92) = 0.14, p > .70$. Means and standard deviations (*SD*) of the prior knowledge sum scores per experimental treatment condition and per study group are provided in Table 7.3.

Table 7.3.
Means (and *SD*) of the prior knowledge scores per experimental treatment condition and per study group

Treatment condition	PBL		GPBL	
Prior knowledge score (0-10)				
	G1	2.20 (1.23)	G3	3.00 (0.87)
	G2	3.56 (1.74)	G4	3.10 (1.10)
	G6	3.63 (1.30)	G5	3.44 (1.67)
	G7	3.30 (1.49)	G8	3.13 (1.25)
	G10	2.82 (1.72)	G9	3.20 (1.75)
Total		3.06 (1.55)		3.17 (1.32)

Sum scores ranged from 0 to 7, and the average sum score of 3.12 indicates that most students did not have much prior knowledge of the topic. Further, the number of misconception alternatives chosen ranged from 1 to 7 and was on average 3.81, indicating that students' low prior knowledge sum score does not reflect mere guessing; it rather reflects that, as to be expected (Ben-Zvi & Garfield, 2004; Garfield, 2003), beginning students have quite some misconceptions about inferential statistics.

7.3.2. Students' perceived value and usefulness of learning statistics

Means and standard deviations of students' perceived value and usefulness scores per experimental treatment condition and per study group are presented in Table 7.4.

Table 7.4.

Means (and *SD*) of students' perceived value and usefulness scores per experimental treatment condition and per study group

Treatment condition	PBL		GPBL	
Perceived value and usefulness score (1-7)				
	G1	5.59 (0.69)	G3	5.73 (0.84)
	G2	5.79 (0.73)	G4	5.59 (0.92)
	G6	5.54 (0.83)	G5	5.73 (1.26)
	G7	5.51 (0.58)	G8	6.20 (0.72)
	G10	4.70 (0.74)	G9	5.90 (0.67)
Total		5.40 (0.79)		5.82 (0.89)

N.B.: the score for a person is an average of scores over seven items in the scale

Unstandardized regression coefficients and corresponding *F*-values and *p*-values for the effects of experimental treatment and students' prior knowledge and interaction between experimental treatment and students' prior knowledge on students' perceived value and usefulness of learning statistics are presented in Table 7.5.

Table 7.5.

Unstandardized regression coefficients and corresponding *F*-values and *p*-values for the effects of experimental treatment and students' prior knowledge and interaction between experimental treatment and students' prior knowledge on students' perceived value and usefulness of learning statistics (i.e., Model 2)

Covariate	<i>B</i>	<i>SE</i>	<i>F</i> (1, 90)	<i>p</i> -value
Intercept	5.40	0.12	1951.05	< .001
Prior knowledge	0.02	0.08	0.08	.780
Treatment *	0.42	0.17	5.75	.019
Treatment by prior knowledge	-0.07	0.12	0.33	.570

* 0 = PBL, 1 = GPBL

As expected, students' perceived value and usefulness of learning statistics was, on average, significantly higher among students in the GPBL groups than in the PBL groups. This finding does not change substantially after removing the non-significant interaction between experimental treatment and students' prior knowledge: the effect of prior knowledge is still not statistically significant, $F(1, 91) = 0.01$, $B < 0.01$, $SE = 0.06$, $p > .90$, while the effect of experimental treatment condition remains statistically significant, $F(1, 91) = 5.78$, $B = 0.42$, $SE = 0.17$, $p = .018$. The latter effect remains unchanged after removal of the non-significant prior knowledge effect, $F(1, 92) = 5.83$, $B = 0.42$, $SE = 0.17$, $p = .018$. The standardized regression coefficient for the experimental treatment effect, $\beta = 0.24$, which indicates a medium size effect.

On the other subscales, as well as on overall IMI score, no significant differences were found between the two treatment conditions. For overall IMI score, $F(1, 92) = 0.22$, $B = 0.05$, $SE = 0.10$, $p > .60$. The same holds for the interaction between experimental treatment and students' prior knowledge: the (non-significant) difference between the two treatment conditions does not depend (significantly) on students' prior knowledge. With regard to students' perceived competence in the domain (one of the seven subscales), a positive prior knowledge effect was found, $F(1, 92) = 5.03$, $B = 0.18$, $SE = 0.08$, $p = .027$. However, after Bonferroni correction for multiple testing, this effect is no longer statistically significant.

7.3.3. Students' conceptual understanding of statistics

To begin with, differences in students' scores between the two test scores (i.e., immediate test performance and delayed test performance) could not be explained significantly in terms of experimental treatment, $F(1, 90) = 0.13$, $B = 0.10$, $SE = 0.28$, $p > .70$, students' prior knowledge score, $F(1, 90) = 0.21$, $B = -0.06$, $SE = 0.13$, $p > .65$, or the interaction between students' experimental treatment and students' prior knowledge, $F(1, 90) = 0.32$, $B = 0.11$, $SE = 0.20$, $p > .55$. Therefore, means and standard deviations of students' conceptual understanding score per experimental treatment condition and per study group as presented in Table 7.6. represent scores averaged over immediate and delayed recall.

The same holds for the unstandardized regression coefficients and corresponding F -values and p -values for the effects of experimental treatment and students' prior knowledge and interaction between experimental treatment and students' prior knowledge on students' conceptual understanding score in Table 7.7.

The interaction effect is statistically significant and the standardized regression coefficient for this effect, $\beta = 0.29$, indicates a medium size effect. A separate analysis for the effect of prior knowledge per experimental treatment condition indicates a non-significant effect in the PBL condition, $F(1, 47) = 0.13$, $B = 0.04$, $SE = 0.12$, $p > .70$, the standardized regression coefficient being $\beta = 0.04$ as well. In the GPBL condition, however, the prior knowledge effect is statistically significant, $F(1, 45) = 17.10$, $B = 0.60$, $SE = 0.14$, $p < .001$, and the standardized regression coefficient $\beta = 0.62$ indicates a large size effect.

The effects reported in Table 7.7. do not change substantially when analyzing the data on immediate and delayed test performance separately, and the interaction between students' experimental treatment and students' prior knowledge remains statistically significant for both immediate recall, $F(1, 90) = 5.68$, $B = 0.50$, $SE = 0.21$, $p = .019$, and delayed recall, $F(1, 90) = 7.88$, $B = 0.61$, $SE = 0.22$, $p = .006$. The positive regression coefficients indicate that once students have more than novice prior knowledge, GPBL enhances conceptual understanding more than PBL.

The standardized regression coefficients, $\beta = 0.24$ for the immediate recall and $\beta = 0.28$ for the delayed recall, indicate medium size effects.

Table 7.6.

Means (and *SD*) of students' conceptual understanding scores per experimental treatment condition and per study group

Treatment condition	PBL		GPBL	
Conceptual understanding score (0-10)				
	G1	5.35 (1.42)	G3	5.17 (1.12)
	G2	4.67 (1.12)	G4	4.85 (1.20)
	G6	5.38 (1.25)	G5	5.28 (0.79)
	G7	5.10 (0.74)	G8	4.69 (1.81)
	G10	5.55 (1.67)	G9	5.60 (2.22)
Total		5.22 (1.27)		5.13 (1.49)

Table 7.7.

Unstandardized regression coefficients and corresponding *F*-values and *p*-values for the effects of experimental treatment and students' prior knowledge and interaction between experimental treatment and students' prior knowledge on students' conceptual understanding score (i.e., Model 1)

Covariate	<i>B</i>	<i>SE</i>	<i>F</i> (1, 90)	<i>p</i> -value
Intercept	5.22	0.19	795.08	< .001
Prior knowledge	0.04	0.12	0.13	.723
Treatment *	-0.12	0.26	0.22	.643
Treatment by prior knowledge	0.55	0.19	8.66	.004

* 0 = PBL, 1 = GPBL

7.3.4. Active participation in the group discussion and students' conceptual understanding

Table 7.8. presents the number of students that participated actively in the group discussion per experimental treatment condition and per study group. In the PBL groups, on average 64.6 percent of the students participated actively in the discussion. In the GPBL, this was 80.4 percent.

Unstandardized regression coefficients and corresponding Wald χ^2 -values and *p*-values for the effects of experimental treatment and students' prior knowledge and interaction between experimental treatment and students' prior knowledge on the likelihood of participating actively in the group discussion are presented in Table 7.9.

Although the main effect of experimental treatment is not statistically significant, the interaction between experimental treatment and students' prior knowledge is statistically significant, and the positive regression coefficient indicates that the non-significant difference between GPBL and PBL in activity during the group discussion for lower levels of prior knowledge increases as prior knowledge increases.

Table 7.8.

Number of students (of total number of students in the group) that participated actively in the group discussion per experimental treatment condition and per study group

Treatment condition	PBL		GPBL	
Number of students (of total number of students in the group)				
	G1	5 (out of 10)	G3	7 (out of 10)
	G2	2 (out of 9)	G4	7 (out of 10)
	G6	6 (out of 8)	G5	8 (out of 9)
	G7	7 (out of 10)	G8	5 (out of 8)
	G10	11 (out of 11)	G9	10 (out of 10)
Overall number	31 (out of 48)		37 (out of 46)	
Total	64.6		80.4	

Table 7.9.

Unstandardized regression coefficients and corresponding Wald χ^2 -values and p -values for the effects of experimental treatment and students' prior knowledge and interaction between experimental treatment and students' prior knowledge on the likelihood of participating actively in the group discussion (i.e., Model 3)

Covariate	B	SE	Wald $\chi^2(1)$	p -value
Intercept	0.60	0.52	1.33	.249
Prior knowledge	-0.36	0.06	0.32	.571
Treatment *	1.08	0.66	2.69	.101
Treatment by prior knowledge	0.86	0.32	7.09	.008

* 0 = PBL, 1 = GPBL

A linear multilevel regression model with students' conceptual understanding as dependent variable and being active or not as covariate reveals that active participation in the group discussion and students' conceptual understanding are correlated, $F(1, 92) = 12.71$, $B = 1.07$, $SE = 0.30$, $p = .001$. Thus, the experimental treatment effect is partly mediated by active participation: GPBL is likely to enhance active participation – and this tendency increases as students' prior knowledge increases – and active participation positively affects students' conceptual understanding of statistics.

The unstandardized regression coefficients and corresponding F -values and p -values for the effects of experimental treatment, students' prior knowledge, the interaction between experimental treatment and students' prior knowledge, and active participation on students' conceptual understanding score are presented in Table 7.10.

The interaction between experimental treatment condition and students' prior knowledge is still statistically significant, and so is the main effect of active participation. The standardized regression coefficients, $\beta = 0.24$ for the interaction effect and $\beta = 0.30$ for active participation, indicate medium size effects.

Table 7.10.

Unstandardized regression coefficients and corresponding *F*-values and *p*-values for the effects of experimental treatment, students' prior knowledge, their interaction effect, and active participation on students' conceptual understanding score (i.e., Model 5)

Covariate	<i>B</i>	<i>SE</i>	<i>F</i> (1, 90)	<i>p</i> -value
Intercept	4.63	0.26	320.51	< .001
Active in group discussion *	0.91	0.29	9.66	.003
Prior knowledge	0.05	0.12	0.19	.662
Treatment **	-0.26	0.26	1.05	.309
Treatment by prior knowledge	0.46	0.18	6.33	.014

* 0 = not active, 1 = active

** 0 = PBL, 1 = GPBL

7.3.5. Learning goals, argumentation, tutors' impressions, and students' opinions

All three tutors indicated that the additional guidance in GPBL positively contributed to the information-richness of and active participation of students in the group discussion. All three tutors indicated that, overall, the GPBL group discussions were more effective and more efficient than the PBL group discussions. However, each of the tutors indicated that – in the statistics knowledge domain – both formats would profit from a more active tutor. In the current study, the difference between the formats was whether or not the learning goals were predefined and the tutor remained more or less in the background.

With regard to the question which of the formats would yield a more enhance knowledge and understanding, the opinions were somewhat divided. One tutor indicated GPBL as preferable format. The other two tutors expect that the predefined learning goals may stimulate specific reproduction rather than overall understanding of the subject. Further, one of the tutors indicated that beginning students may miss the reasoning and argumentation skills necessary to formulate good arguments (as in MPM) by means of their answers to predefined learning goals in GPBL. This indication is in line with content analysis of the GPBL group discussions, from which becomes clear that in only two of five GPBL conditions the argumentation went more or less fluently and in one other group, the argumentation appeared to not work at all.

However, the tutors share the expectation that beginning students miss knowledge and skills to formulate good learning goals and find their way in the statistics knowledge domain by means of learning goals that are the product from group discussion. Again, this expectation is in line with content analysis of the PBL group discussions. Although some of the learning goals, formulated in three of the five PBL groups, would help students in their study (e.g., "what is a null hypothesis?", "what is a sampling distribution?"), other learning goals were of a lower level than addressed by the problem and literature to be studied (e.g., in one group: "what is a median?", and in other groups: "what is a mean?" and "what is a standard deviation?"), and in each of the five PBL groups, the learning goals together did not cover the topic to be studied.

The majority of students of both the PBL and GPBL groups indicated that they felt they could have learned more from the group discussion, had a lecture preceded the group discussion and had the tutor been more active. The latter is in line with the tutors' impressions.

7.4. Discussion

We hypothesized that GPBL would enhance both students' conceptual understanding of statistics and students' perceived value and usefulness of learning statistics. The results indicate that presenting students with predefined learning goals as in GPBL indeed tends to raise students' awareness of the value and usefulness of learning statistics. Although even in the PBL condition the average score on this scale was fairly high (5.40), the increase in average score of 0.42 points in the GPBL condition reflects a medium size effect. With regard to conceptual understanding, the results indicate that, on average, GPBL tends to enhance conceptual understanding of statistics more than PBL once students have some prior knowledge. The effect of prior knowledge on conceptual understanding within the PBL condition is non-significant. In the GPBL condition, the effect of prior knowledge on conceptual understanding is large ($\beta = 0.62$).

7.4.1. An implication for the teaching practice

Taken together, the findings suggest that to raise students' awareness of why learning statistics is important and why the exercises they are doing in the group are useful, and to raise students' conceptual understanding of statistics, predefined learning goals should be used to structure group discussions.

The average prior knowledge in the current study was very low (3.12 on a scale ranging from 0 to 10). Only a small proportion of students had obtained some knowledge about hypothesis testing and sampling distributions in their high school mathematics course, and even these students indicated that they had forgotten most of it again. An explanation for the latter is that high school mathematics courses focus on computational aptitude or the ability to understand formulas and to use them correctly (Huberty et al., 1993), whereas in university curricula for the social and health sciences the focus is on conceptual understanding of statistics (Broers, 2002, 2009). Therefore, to ensure that beginning students in these curricula have sufficient prior knowledge to start their learning journey in the statistics knowledge domain, an introductory lecture should precede the group discussion. The lecture can raise prior knowledge as well as awareness of the value and usefulness of learning statistics among students. If instructors then choose for a guided form of PBL, namely GPBL, this can (1) increase students' awareness of the value and usefulness of learning statistics further, (2) increase the proportion of students that actively participates in the group discussion, and (3) enhance conceptual understanding more than the classical PBL setup.

7.4.2. Suggestions for further research

Taking into consideration tutors' and students' notion that a more active tutor could improve the effectiveness of the group discussions – a notion that is in line with findings in previous studies (Bude et al., 2009; Chang, Yew, & Schmidt, 2011) – teaching practice should consider a more active role of the tutor. This is also an interesting question for future research. In the current study, the role of the tutor was rather passive, (s)he only stimulated knowledge elaboration by asking questions whenever the group discussion led to a dead end or to misconceptions, (s)he did not share any knowledge with the students. A new study could compare PBL and GPBL in terms of conceptual understanding in a condition of more active tutors. Tutor activity (active or rather passive) could serve as additional factor, to examine whether differences between PBL and GPBL in terms of conceptual understanding or perceive value and usefulness of learning statistics

depends on the level of tutor activity. Besides, a more active tutor could lead to an increase in the proportion of students that participates actively in the group discussion and it could contribute to other aspects of students' motivation to learn statistics (e.g., a feeling of trust and social support from the tutor). Thus, tutor activity is an interesting factor for future research for a variety of reasons. To examine the separate effects of tutor presence and activity and group discussion (guided or unguided), distinct types of control conditions could be considered. If in a study, the learning goals that are provided to GPBL prior to self-study are closely aligned with the conceptual understanding test questions, it is possible that either group discussion or tutor presence are not critically important to demonstrate gains for GPBL over PBL. A control group that receives learning goals and self-study, but no group discussion, could then be important to evaluate this possibility. To evaluate in this context the added value of tutor presence and/or tutor activity, this control group could be compared to groups in which group discussion is led by a tutor – who is then either active or not – and a group in which the group discussion is done without a tutor.

A second important factor for future research is the nature of the subject matter within the statistics knowledge domain. The subject matter in our experiment was sampling distributions and null hypothesis testing, and the students who participated were university freshmen who mostly had limited prior knowledge. As an understanding of the theory around sampling distributions and null hypothesis testing is essential for further successful study in the statistics knowledge domain – this subject matter is the very fundament of most of statistics students are taught later on in their curriculum – an interesting question is how effective PBL and GPBL are for more advanced statistical topics, topics that require (prior) knowledge of sampling distributions, null hypothesis testing, and related concepts.

A third factor of interest may be the knowledge domain itself. As mentioned previously, the extent of instructional guidance needed is likely to depend on students' prior knowledge as well as on the complexity of the knowledge domain. An interesting comparison would be how effective PBL and GPBL are in different knowledge domains, for example statistics versus medicine, or statistics versus social psychology.

Further, two recent experiments suggest that presenting students with fully worked-out examples (Leppink et al., 2011b; Chapter 4 of this thesis) or partially worked-out examples (Leppink et al., 2012; Chapter 5 of this thesis) of GPBL-like learning goals may help to enhance conceptual understanding among students who have very limited prior knowledge. However, these two experiments focused on individual students and students learning in pairs. An interesting question for future research is therefore whether discussing partially or fully worked-out examples in a GPBL group can enhance conceptual understanding among students who have little to no prior knowledge as well. The current experiment demonstrates that some prior knowledge is a prerequisite for a positive effect of GPBL relative to PBL.

The experiment described in this article may contribute to a new line of research of dozens of potential experiments varying in experimental manipulations. Some of these manipulations (e.g., tutor activity) may change some of the effects demonstrated in the current experiment. Through interaction of theory and empirical research we may be able to find optimal, possibly more guided, PBL formats for different knowledge domains, including the domain of statistics.

Chapter 8

Discussion

This thesis started from the observation that statistics is considered a difficult subject and that, for a number of reasons, many students develop only superficial understanding of this subject. Appropriate teaching methods are needed to help students develop knowledge and understanding of statistics from the very start and step by step.

8.1. How the chapters combine to form a line of research

The studies that form the core of this thesis were carried out to examine the potential of an instructional method, called the method of propositional manipulation (MPM; Broers, 2002; 2008), that was developed for the statistics knowledge domain, in individual learning settings and in group learning environments. Although students usually develop some knowledge of statistical concepts and ideas (i.e., propositional knowledge), many of them find it difficult to understand how these concepts and ideas are interrelated. The latter is called conceptual understanding. The main research question throughout the research project was whether MPM can help students build conceptual understanding of statistics.

From Chapters 2-5 it becomes clear that in individual learning settings, sufficient prior knowledge is a necessary condition for MPM being an effective format to help students develop conceptual understanding of statistics. The more knowledgeable students tend to profit from the individual performance of a series of MPM learning tasks (a finding throughout Chapters 2-5), whereas students who have very limited prior knowledge tend to profit more from self-explaining specific passages of a study text (Chapter 3), from studying fully worked-out examples of MPM learning tasks (Chapter 4) and from studying partially worked-out examples individually, or even better: in pairs of students (Chapter 5).

Chapter 6 suggests that in an interactive lecture setting having more prior knowledge is no longer a necessary condition for a beneficial effect of MPM on students' conceptual understanding: students who have very little prior knowledge profit equally well from MPM as more knowledgeable students do. However, in line with Chapters 2-5, Chapter 7 demonstrates that – when applied as a method to structure group discussion – the more knowledgeable students develop significantly more conceptual understanding than students who have little to no prior knowledge.

Although each of the Chapters 2-7 focused on a different research question, these research questions were interrelated and all focused on the main question whether MPM can help students build conceptual understanding of statistics. Taken the findings of Chapters 2-7 together, the question can be answered as follows: in its original setup (i.e., having students perform MPM learning tasks themselves, not in a fully or partially worked-out examples setup), MPM can help the more knowledgeable students enhance conceptual understanding of statistics even in individual learning settings, whereas the less knowledgeable students need additional instructional guidance from an instructor (i.e., lecturer or knowledgeable tutor) to develop conceptual understanding of the domain via MPM. If such instructional guidance is not considered an option by the instructors or curriculum developers, fully or partially worked-out

examples – together with the instruction to work in pairs or groups of students – should be considered.

Not only do findings from our empirical research provide us with answers to (some of) our research questions – and hopefully give practical implications as well – they also lead to new questions and suggestions for new research. The findings reported in Chapters 2-7 build on each other. As is the case for many doctoral theses, the sum is bigger than its parts; the implications and suggestions that follow from Chapters 2-7 together have more power than the implications and suggestions that follow from any of the separate chapters. In the remainder of this chapter, first of all, implications and suggestions that follow from the previous chapters are discussed in a theoretical framework. Subsequently, the limitations of the studies presented in the previous chapters are discussed, and these limitations give rise to suggestions for new studies.

8.2. Knowledge elaboration in individual learning groups

The interaction between the levels of students' prior knowledge and the effectiveness of different instructional formats is called the expertise reversal effect (Kalyuga, 2005, 2006, 2007; Kalyuga et al., 2001a, 2003). It is a robust finding that has now also been demonstrated in the statistics knowledge domain (Chapters 3-5 of this thesis): only students who have sufficient prior knowledge of the domain can build conceptual understanding by performing MPM learning tasks individually.

As mentioned in Chapter 1, the two most successful cognitive mechanisms of self-explanations appear to be filling knowledge gaps (Chi, 2000) and constructing knowledge networks (Novak, 2002). Studies on the expertise reversal effect are intuitively appealing in that they appear to demonstrate that less knowledgeable students need to fill their knowledge gaps (i.e., through the study of worked-out examples), while more knowledgeable students have sufficient knowledge to construct (and enhance) knowledge networks through argumentation. Which component of self-explanation is effective for an individual student appears to depend largely on the student's prior knowledge of the subject. Therefore, to be a potentially successful format for students who lack prior knowledge as well, MPM learning tasks need to be presented in the form of either fully worked-out examples (Chapter 4) or partially worked-out examples that can be studied in pairs of students (Chapter 5). Possibly, a stage of MPM learning tasks presented in the form of (fully or partially) worked-out examples could precede a stage in which students perform MPM learning tasks in its original setup individually.

8.3. Knowledge elaboration in group learning environments

The expertise reversal effect has been studied more extensively in individual learning settings rather than in group learning environments. Although, two studies found that potentially beneficial effects of knowledge elaboration can be moderated by students' prior knowledge in group learning environments as well (Willoughby et al., 1993; Woloshyn et al., 1992). On the one hand, students who have insufficient prior knowledge may not be able to engage in knowledge elaboration. On the other hand, sufficient prior knowledge may provide students with a contextual framework for new and autonomous learning (Machiels-Bongaerts, Schmidt, & Boshuizen, 1995; Peeck, 1982; Pichert & Anderson, 1977; Wetzels, Kester, & Van Merriënboer, 2011).

Chapter 5 focused on the difference between passive learning, active learning, constructive learning, and interactive learning. Self-explanation is a potentially effective constructive learning process, while explanation to peers is a potentially effective interactive learning process. Collaborative (group) learning can promote active learning (O'Donnell, 2006) as well as constructive and interactive learning (Chi, 2009; Fonseca & Chi, 2011). Assuming that two students know more than one, additional (prior) knowledge is likely to become available by means of collaborative learning (Johnson et al., 2007). Providing explanations to others can help students to enhance their own knowledge structures (Krol, Janssen, Veenman, & Van der Linden, 2004; Nussbaum, 2008; Webb, 1989) and is therefore assumed to be an important mechanism in the effectiveness of collaborative learning. Van Blankenstein, Dolmans, Van der Vleuten, and Schmidt (2009) demonstrated that providing explanations during a group discussion may enhance students' knowledge and understanding more than merely listening. However, this may hold only in the case students have sufficient prior knowledge. In the latter study, for example, prior to participation in the study, students were given quite extensive training in the topic.

Chapters 6 and 7 focused on the effect of guidance in group learning for the statistics knowledge domain. A crucial difference in setup between Chapters 6 and 7 may be that in Chapter 6 the discussion was led by an active lecturer, whereas the tutor in Chapter 7 remained relatively on the background and only stimulated knowledge elaboration via standard questions in the tutor protocol. In the light of the findings presented in Chapter 6, and in line with previous research (Bude et al., 2009; Chang et al., 2011), it is likely that the findings reported in Chapter 7 change more positively if the tutor becomes more active and shares some of his or her knowledge with the students in the group.

8.4. Instructional guidance should decrease with increasing prior knowledge

Having new students to learn by themselves in the domain of statistics may quickly lead to disorientation on the part of the students. There is a need for instructional formats that stimulate students to self-explain without experiencing disorientation. The constructivist view that knowledge structures have to be actively constructed and self-explained by the student (Novak, 1998) is likely to be correct. However, students should be guided in this process to avoid disorientation. This has been the focus of the original MPM format. MPM is intended to guide students to handle and understand different ways to solve a problem and to develop problem solving strategies. Together with more partially and fully worked-out variants, the MPM format provides a way to deal with some of the educational factors as well as with some of the student-related factors that make teaching and learning statistics so challenging. Students need to relate and integrate concepts and propositions to form and develop schemata of the subject matter. Flexible guidance, that is: students' prior knowledge level determines the extent of guidance needed (i.e., fully or partially worked-out examples on the one hand, formulating arguments on the other hand), aims to stimulate students to engage in meaningful learning, fill important knowledge gaps, and gradually develop knowledge networks that integrate statistical concepts and ideas.

Together, the six empirical studies presented in this thesis suggest that instructional guidance should decrease as students' prior knowledge increases. To have students learn a new topic, the following teaching strategy could be adopted:

- *Stage I, When most students have little prior knowledge:* the topic is introduced in an interactive lecture, and examples of MPM learning tasks are presented and/or discussed; students' prior knowledge of the topic is increased by having them fill crucial gaps in their knowledge.
- *Stage II, The interactive lecture is followed by group learning:* the topic is discussed (pre self-study), self-studied, and discussed again (post self-study) in small guided problem-based learning (GPBL) groups; possibly, (partially or fully) worked-out examples of MPM learning tasks can be discussed in the discussion prior to self-study, and actual performance of MPM learning tasks can be the main activity in the self-study and discussion after self-study; a knowledgeable tutor can provide flexible guidance. The fundament for the construction and enhancement of knowledge networks is created.
- *Stage III, Individual practice of argumentation:* students should now have sufficient prior knowledge to perform autonomously a series of MPM learning tasks by bearing in mind a limited set of rules, propositions, or premises, which is accomplished in the MPM format by having students integrate a number of manipulated propositions into an argument. Students should be given the opportunity to acquire feedback on their task performance (e.g., response lecture).
- *Stage IV, Demonstrating conceptual understanding:* as in real-life practice or research, one needs to select relevant rules, propositions, or premises themselves, and apply knowledge and understanding autonomously – without (instructional) guidance – problem-solving in this phase should be individually and unguided. In this stage, instead of confronting students with an MPM learning task, one could just present the hypothesis and instruct students to select relevant propositions themselves, and integrate these propositions into an argument that proves the hypothesis true or false.

Depending on the course aims, a combination of the *Stage III* and *Stage IV* problem-solving approach could be applied in exam situations (e.g., two different types of problems). When applying the unguided approach in an exam following course stages that apply the guided and self-guided approach, an interesting question is how students solve the problems in an exam, more specifically: will students still solve the problems by integrating relevant propositions into arguments? If yes, they demonstrate conceptual understanding, and that they can apply (some of) their knowledge and communicate with others in a way that makes sense. Further, it has been found that combining an instructional method and an exam situation enhances learning more than an instructional method without exam situation (Karpicke & Roediger, 2008; Larsen, Butler, & Roediger, 2008).

Although these four stages may make sense to many curriculum developers, instructors, and students, the educational practice is different. If lectures are part of a course, they are rarely interactive and the slides present many formulas, graphs, and tables without much explanation. Even if the course starts in *Stage I* or *II*, the subject materials in lectures or group learning settings is commonly presented without much instructional guidance. In traditional problem-based learning (PBL) groups, for example, students are supposed to formulate learning goals by means of group discussion with their peers. Chapter 7 of this thesis demonstrates that such learning goals tend to be very broad and of a complexity level that is considerably lower than the level they are supposed to be. In the MPM format, it is the instructors who set the complexity

level and this complexity level should be in line with students' prior knowledge of the subject. If the learning goals formulated by the instructors are too easy, students will be bored and consider studying them a waste of time; if the learning goals formulated by the instructors are too complex, the less knowledgeable students are likely to get lost.

8.5. Strengths and limitations

As any scientific work, the studies presented in this thesis have strengths and limitations that deserve critical evaluation, and new studies should take at least some of the limitations into account.

The strengths of the studies presented in this thesis are: the mixed method approach (especially in Chapter 2 and to a lesser extent in Chapter 3), the fact that five of the six studies are randomized experiments in a real education context or similar context (Chapters 3-7), that the interaction between students' prior knowledge and instructional formats has been demonstrated in different experiments (Chapters 4, 5, and 7, and to a lesser extent also in Chapter 3), and students were sampled from different cohorts.

To explore how an instructional format like MPM works in practice and to acquire an overall idea about what factors influence a student's ability to learn from MPM learning tasks and how these factors interact, it is important to combine different research methods. The study presented in Chapter 2 therefore combined quantitative measures for cognitive load and qualitative measures (i.e., a mixed method approach), and used a technique from the cognitive research tradition, namely thinking aloud while performing a series of MPM learning tasks.

The study presented in Chapter 2 provided us with insights in how to design the randomized experiments presented in Chapters 3-7. These five chapters demonstrate – contrary to what some may believe – that randomized experiments are feasible in educational research and should be preferred over quasi-experimental and observational designs.

Although sample sizes in randomized experiments are generally smaller than sample sizes in quasi-experimental and observational studies, causal inference based on randomized experiments is generally stronger than causal inference based on quasi-experimental and observational studies. Some may state it even stronger, namely that it is arguably impossible to detect true causal effects from the latter, as in the latter internal validity is bound to be limited. In quasi-experimental and observational studies unbalance in the design and various other confounding variables may resonate in the effects found (and we can rarely correct for all confounding variables). Confounding variables not taken into account leave room for alternative explanations for treatment effects found and undermine internal validity of findings. Randomized experiments are a powerful tool to rule out alternative explanations for treatment effects, and stratified random sampling (like in Chapters 4, 5 and 7) can help to minimize unbalance in the design.

Advocates of quasi-experimental and observational designs may criticize that external validity of the findings from randomized experiments is limited, for randomized experiments may be conducted in a context that is isolated from real education. However, Chapters 5 and 6 present two experiments in which the participating students were preparing for their re-sit exam, and Chapter 4 and 7 present two experiments in which the participating students were about to start their first statistics course in university. Further, in all the experiments presented in Chapters 3-7, the materials students had to study were directly relevant for their studies. Students who

participated in one of the experiments presented in Chapters 4-7 had a stake in their performance and received feedback on their performance (i.e., test results) after participation. In the experiments in Chapters 4 and 7, feedback was provided in individual feedback meetings. In the experiments in Chapters 5 and 6, students attended one or more lectures on the study materials after participation and had the opportunity to have individual or small group feedback meetings. Finally, the fact that for each of the studies presented in this thesis students were sampled from different cohorts and that the interaction between students' prior knowledge and instructional formats has been demonstrated in different experiments adds up to the generalizability of findings. The expertise reversal effect is a robust finding that has important implications for statistics education (see the four stages in paragraph 8.4.). Curriculum developers and teachers should inform themselves about these implications and present their learning materials accordingly. To conclude, we believe that the randomized experiments presented in this thesis do not suffer that much from limited external validity.

A first limitation of the studies presented in this thesis, and especially the studies in Chapters 2-3 – is that they are based on rather small sample sizes. As a consequence, small to medium size effects may not have been detectable by means of statistical significance testing, and especially tests for interaction and simple effects may have suffered from limited statistical power. Also, the effects in each of the studies depend on the limited size and content of the learning materials used. It is very well possible that the use of other learning materials would lead to smaller or larger effects compared to the ones presented in this thesis. Also, the studies presented in this thesis largely focused on basic inferential statistics (Chapters 2-5 and 7), and to a lesser extent on slightly more advanced inferential statistics (Chapter 6) and descriptive statistics (Chapter 2). In most of the studies, the subject matter was sampling distributions and null hypothesis testing, and the students who participated were in most cases university freshmen who mostly had limited prior knowledge. An understanding of the theory around sampling distributions and null hypothesis testing is essential for further study in the statistics knowledge domain. This subject matter is the very fundament of most statistical methods that are taught later on in many curricula (e.g., linear regression) and of methods that are part in still few curricula nowadays (e.g., Bayesian methods). An interesting question is how effective MPM learning tasks and (fully or partially) worked-out examples of such learning tasks are for more advanced statistical topics, topics that require (prior) knowledge of sampling distributions, null hypothesis testing, and related concepts. In sum, new studies should use sufficiently large sample size, use different learning materials, and focus on more advanced topics within the statistics knowledge domain.

A second limitation of this thesis is the conceptualization of cognitive load. Although in Cognitive Load Theory, the distinction between the three additive types called intrinsic load, extraneous load, and germane load has been made, valid instruments for the measurement of these types of load are still lacking. In line with research tradition, in the studies presented in Chapters 2-5, one item that is supposed to measure total cognitive load was used. We sought to distinguish between the types of load by administering the one item for total load in both learning stage and testing stage and make a comparison between students' total load ratings and their test performances. However, one single item may be very sensitive to various influences (other than experimental manipulation) which we did not intend to measure. As was argued in Chapters 1-5, for educational purposes the distinction between three types of load is relevant and interesting. The development of instruments that can measure each of these types is

therefore needed and here is a challenge for a new line of research. In this line of research, the study of participants' subjective ratings of items and advanced techniques like functional magnetic resonance imaging (fMRI) can be combined.

A third limitation of this thesis may lie in the coding of propositional knowledge (Chapters 3-5) and conceptual understanding (Chapters 3-7). The items used to measure these response variables originated from real statistics courses, and like in these courses, every item had the same weight. For example, if a test counted ten multiple choice questions, every correctly performed item added 1 point to the student's score while an incorrect response did not add anything. Equal weight for all items may not be defensible. Alternative weighting could occur according to item difficulty level, the extent to which items discriminate between students who do or do not have a certain level of knowledge or understanding, the number of alternatives in an item (which was equal for all items in each of the studies presented), or according to a combination of these item characteristics. However, to determine such item weights, these items need to be subjected to a large sample of participants who have not just experienced different experimental treatments. Future studies are therefore needed, to first subject a pool or set of items to a large group of students, determine a weighting algorithm via item response models, and use that algorithm throughout a series of subsequent randomized experiments like the ones presented in this thesis. It is possible that different weighting algorithms yield somewhat different effects.

A fourth limitation is that each of the studies presented in this thesis suffered from restriction of range problems with regard to students' prior knowledge. As a result, some of the effects reported (and perhaps interaction effects in particular) may be smaller than in real courses. In Chapter 2, the inclusion criterion for participation was that students passed their first inferential statistics course. As a result, participating students mainly represented the upper range of the proficiency scale, and this may explain why students were not necessarily motivated to formulate extensive arguments on a problem of which the solution was clear to them anyway. In Chapters 3-4 and 7, students were university freshmen who had not yet attended any statistics courses in their curriculum. For experimentation purposes, study sessions in the studies presented in these three chapters were much shorter than study sessions in real statistics courses. Besides, most students in real courses attend lectures, read considerably more literature than the four pages presented to them in our studies, and consequently have more time than they had in the study to digest the subject matter. Finally, participants in the studies presented in Chapters 5-6 represented the lower range of the proficiency scale as they were preparing for their re-sit. It is well known that effects in a sample can be somewhat small due to restriction of range. However, the studies reported in Chapters 4-7 demonstrate that randomized experiments can be conducted in a real statistics course or at least in a context that is fairly close to such a course. When students realize that their participation in a study can boost their knowledge and understanding for their exam or another purpose, they will be more inclined to participate than in a context that isolated from such purposes. Most likely for that reason, the studies presented in Chapters 2-3 are based on only small sample sizes.

The critical reader may argue at this point that a fifth limitation of the studies presented in this thesis is that they suffer from a volunteer bias. Although this may be true to some extent, the findings and implications presented in this thesis may apply only to students who are motivated enough to visit lectures, group discussions, and spend sufficient time on studying and exercising.

Students who do not attend any lectures or group discussions and who spend a minimum amount of time in studying (and eventually skip the exercises) are more likely to fail the exam than their more motivated peers, and to such students none of the findings (Chapters 2-7) and implications (in the paragraphs following) may apply.

Finally, the fact that the lecturer in Chapter 6 was active and the tutor in Chapter 7 rather inactive may explain to some extent why the significant interaction effect between experimental treatment and students' prior knowledge found in Chapter 7 was not found in Chapter 6. Further, increased tutor activity may influence students' active participation in group learning environments (Bude et al., 2009; Chang et al., 2011). Different opinions about the role of the tutor in learning environments exist. In Cambridge, for example, every student at a College has a personal tutor who is not only knowledgeable in the learning domain but also keeps an eye on the student's social well-being. Tutors are fellows or teachers who have found their way in the learning domain and guide, assess, and advise small numbers of students. Tutor activity may influence students' motivation to learn, students' social well-being in learning environments, students' participation in these learning environments, and their development of conceptual understanding of a complex knowledge domain like statistics.

English summary

Six empirical studies centered around one main research question: can the method of propositional manipulation (MPM) help students build conceptual understanding of statistics? In each study, MPM was compared and contrasted with one or more alternative instructional formats. The first three studies focused on the student as individual learner, the last three studies focused on group learning rather than individual learning. **Chapter 1** presents the theoretical framework for this thesis as well as an overview of the distinct research questions covered by **Chapters 2-7**. **Chapter 2** explores the main factors affecting a student's ability to learn from performing MPM learning tasks, by having twenty undergraduate psychology students perform six MPM learning tasks while thinking aloud. The results indicate that whether students learn from MPM depends on their statistics proficiency level, the subject matter, the number of propositions in the learning task, and the instructions. MPM learning tasks should be tailored to the students' level of expertise and students should be instructed more than once to integrate all propositions in the learning task into their arguments. **Chapters 3-7** present the findings of a total of five randomized controlled experiments. **Chapter 3** presents the findings of a randomized controlled experiment in which two self-explanation conditions – one guided and one unguided – and a reading (control) condition were compared to examine the effect of guiding self-explanation on cognitive load, propositional knowledge, and conceptual understanding of statistics. The results indicate that students self-explaining the subject matter when studying experience a lower cognitive load during exam situations. Moreover, guiding students into self-explanation appears to enhance conceptual understanding more than unguided self-explanation only for students who have already become aware of most of their misconceptions.

Chapter 4 presents the findings of a randomized controlled experiment on the effects of four instructional methods on cognitive load, propositional knowledge, and conceptual understanding of statistics, for low prior knowledge and for high prior knowledge students. The instructional methods were a reading-only control condition, answering open-ended questions, answering open-ended questions and formulating arguments, and studying worked-out examples of the type of arguments students in the third group had to formulate themselves. The results indicate that high prior knowledge students develop more propositional knowledge of statistics than low prior knowledge students. With regard to conceptual understanding, the results indicate an expertise reversal effect: low prior knowledge students profit most from worked-out examples, while high prior knowledge students profit most from formulating arguments. Thus, novice students should be guided into the subject matter by means of worked-out examples. As soon as students have developed more knowledge of the subject matter, they should be provided with learning tasks that stimulate students to solve problems by formulating arguments.

Chapter 5 presents the findings of a randomized controlled experiment on the effects on different teaching and learning methods for statistics for two levels of prior knowledge on cognitive load, propositional knowledge, and conceptual understanding of statistics. Teaching methods were whether or not to provide students with propositional information and learning strategies were self-explaining the learning material and explaining in pairs. The results indicate that prior knowledge facilitates propositional knowledge development and leads to differential effects of teaching and learning methods on conceptual understanding: only low prior knowledge

students profit from additional information in the learning task and/or explaining in pairs. An implication of these findings is that low prior knowledge students should be guided into the subject matter by means of working in pairs on learning tasks that comprise additional information. Once students have developed more knowledge of the subject matter, they should be stimulated to work individually on learning tasks that do not comprise additional information.

Chapter 6 presents the findings of a randomized controlled experiment on the potential effects of MPM as a lecturing method on motivation to learn and conceptual understanding of statistics. MPM aims to help students develop conceptual understanding by guiding them into self-explanation at two different stages. First, at the stage of propositions (statements referring to single statistical concepts and ideas), and subsequently, at the stage of more complex problems that comprise a set of relevant propositions. A total of 71 bachelor students in psychology who were preparing for the re-sit of their inferential statistics exam participated in one of two possible lectures. Topic, content, lecturer, and duration of both lectures were the same, and in both lectures five true/false hypotheses were presented. Students in the first lecture (control group) discussed interactively the truth or falsity of each hypothesis. In the second lecture (MPM group), this interactive discussion was structured by presenting a number of short open-ended questions along with each hypothesis. Conceptual understanding was measured by means of a twelve items multiple choice test. Further, the intrinsic motivation inventory (IMI) was administered to examine motivation to learn. The results indicate that MPM does not lead to enhanced motivation to learn but can facilitate conceptual understanding development among students.

Chapter 7 presents the findings of a randomized controlled experiment on the effects of problem-based learning (PBL) and a guided form of PBL (GPBL), for different prior knowledge levels, on motivation to learn and conceptual understanding of statistics. Contrary to classical PBL, in GPBL, the learning goals are not a product of a student group discussion; they are determined by the instructor beforehand, based on a detailed decomposition of the subject matter. These learning goals can then be used by the GPBL group to activate prior knowledge and to structure self-study and subsequent group discussion. The participating students were allocated randomly to either standard PBL or GPBL. The results indicate that, on average, GPBL tends to enhance conceptual understanding of statistics more than PBL once students have some prior knowledge of the subject. Further, GPBL tends to raise students' awareness of the value and usefulness of learning statistics. If PBL groups are preceded by lectures in which the subject matter is introduced – and in which students' prior knowledge of the subject can be enhanced – an implication of this study for the teaching practice is that, for the statistics knowledge domain, instructors should consider a more guided form of PBL rather than classical PBL.

Chapter 8 provides a strategy for teaching and learning statistics in a real course. Instead of lobbying for differential education for students from different prior knowledge groups, the key is to decrease instructional guidance as students' prior knowledge increases. Every topic should be introduced in an interactive lecture in which MPM learning tasks are presented and/or discussed, to then have students study (worked-out examples of) MPM learning tasks in (GPBL) groups and, subsequently, individually. In the interactive lecture and GPBL discussions, knowledgeable students can enhance their knowledge networks by explaining to their less knowledgeable peers who then have opportunities to fill important knowledge gaps. The latter can start enhancing their knowledge networks in a later stage of the course.

Nederlandse samenvatting

Zes empirische studies werden uitgevoerd rondom de hoofdvraag of de methode van propositionele manipulatie (MPM) studenten kan helpen bij het ontwikkelen van conceptueel begrip van de statistiek. In iedere studie werd MPM vergeleken met een of meerdere alternatieve instructiemethoden. De eerste drie studies richtten zich op de individuele student, terwijl de laatste drie studies meer op groepsleren gericht waren. In **hoofdstuk 1** wordt het theoretisch kader voor deze dissertatie uiteengezet, alsmede een overzicht van de verschillende onderzoeksvragen die in de **hoofdstukken 2-7** aan bod komen. In **hoofdstuk 2** wordt gerapporteerd over een studie naar factoren die van invloed zijn op het vermogen van een student om MPM-leertaken uit te voeren. Twintig bachelorstudenten in de psychologie voerden hardop denkend een zestal MPM-leertaken uit. De resultaten laten zien dat voorkennisniveau, studiestof, de instructies en het aantal proposities in een MPM-leertaak van invloed zijn op de mate waarin een student de leertaak kan uitvoeren. Een belangrijke implicatie van de studie is dat MPM leertaken toegesneden op het voorkennisniveau van studenten optimaal effect hebben. Bovendien dient de instructie aan studenten om alle proposities in de leertaak te integreren te worden herhaald. In **hoofdstukken 3-7** worden de bevindingen van een vijftal gerandomiseerde experimenten gepresenteerd. In **hoofdstuk 3** wordt een gerandomiseerd experiment gepresenteerd waarin drie condities – alleen lezen, ongestuurde zelfverklaring en gestuurde zelfverklaring – werden vergeleken in termen van cognitieve belasting, propositiekennis en conceptueel begrip van de statistiek. Studenten die tijdens de studiefase aan zelfverklaring doen minder cognitieve belasting ervaren tijdens een examen. Bovendien lijkt voor de studenten die al in een vroeg stadium bewust zijn van belangrijke misconcepties gestuurde zelfverklaring meer bij te dragen aan conceptueel begrip dan ongestuurde zelfverklaring.

In het gerandomiseerd experiment beschreven in **hoofdstuk 4** werden de effecten van vier instructiemethoden vergeleken in termen van cognitieve belasting, propositiekennis en conceptueel begrip, voor studenten met relatief veel voorkennis en studenten met relatief weinig voorkennis. De instructiemethoden waren: alleen lezen (controlegroep), het beantwoorden van open vragen, het beantwoorden van open vragen gevolgd door het formuleren van argumenten en het bestuderen van uitgewerkte voorbeelden van het type argumenten dat studenten in de derde conditie zelf moesten formuleren. De resultaten wijzen uit dat studenten met relatief veel voorkennis meer propositiekennis ontwikkelen dan studenten met relatief weinig voorkennis. Met betrekking tot conceptueel begrip werd een zogeheten *expertise reversal effect* aangetoond: studenten die relatief weinig voorkennis hebben profiteren het meest van uitgewerkte voorbeelden, terwijl studenten die relatief veel voorkennis hebben het meest profiteren van het zelf formuleren van argumenten. Men dient novieten aan de hand van uitgewerkte voorbeelden met de stof bekend te maken. Zodra studenten voldoende kennis van de studiestof hebben, kunnen zij worden gestimuleerd tot het uitvoeren van leertaken waarin het formuleren van argumenten centraal staat.

In de studie gepresenteerd in **hoofdstuk 5** werden de effecten van twee instructiemethoden en twee studiemethoden voor de statistiek onderzocht in termen van cognitieve belasting, propositiekennis en conceptueel begrip, voor studenten met relatief veel voorkennis en studenten met relatief weinig voorkennis. De instructiemethoden waren het uitvoeren van MPM-

leertaken en het uitvoeren van dergelijke leertaken waarvan een deel al was uitgewerkt. De studiemethoden waren het individueel werken en het werken in paren van studenten. Meer voorkennis samengaat met meer propositiekennis na de studiefase en alleen de studenten met relatief weinig voorkennis profiteren van uitwerkingen in leertaken en/of van het werken in paren. Studenten met weinig voorkennis hebben dus sturing nodig in de zin dat zij aan gedeeltelijk uitgewerkte leertaken werken in plaats van aan geheel zelf uit te voeren leertaken en dat daarnaast samenwerking met een medestudent bevorderlijk kan zijn voor de ontwikkeling van conceptueel begrip onder deze studenten. Zodra studenten meer kennis van de studiestof hebben, kunnen zij worden gestimuleerd om individueel aan leertaken te werken en zo ook aan leertaken die geen uitwerkingen bevatten.

In **hoofdstuk 6** worden de bevindingen van een gerandomiseerd experiment gepresenteerd over de mogelijke effecten van MPM als collegemethode op leermotivatie en conceptueel begrip van de statistiek onder studenten. MPM poogt studenten te helpen bij de ontwikkeling van conceptueel begrip door hen te sturen in zelfverklaring op twee verschillende niveaus, allereerst op het niveau van proposities (stellingen die verwijzen naar statistische concepten en ideeën) en vervolgens op het niveau van meer complexe problemen waarin meerdere proposities samenkomen. In totaal 71 bachelorstudenten in de psychologie die bezig waren met de voorbereidingen voor een herkansing namen deel aan het experiment. Via het lot werden zij toegewezen aan een van beide colleges. Onderwerp, inhoud, collegegever en duur waren voor beide colleges hetzelfde en in beide colleges werden vijf juist/onjuist stellingen gepresenteerd. Studenten in het ene college (controlegroep) discussieerden interactief over de juistheid dan wel onjuistheid van iedere stelling. In het andere college (MPM-groep) werd over dezelfde stellingen gediscussieerd, met als enige verschil dat nu iedere stelling samen met een aantal open vragen werd gepresenteerd. Conceptueel begrip werd gemeten aan de hand van twaalf meerkeuzevragen en leermotivatie werd gemeten aan de hand van de zogeheten *intrinsic motivation inventory* (IMI). The resultaten laten zien dat MPM niet bijdraagt aan meer leermotivatie maar wel een gunstig effect op de ontwikkeling van conceptueel begrip kan hebben.

In **hoofdstuk 7** gaat het om een gerandomiseerd experiment waarin probleem gestuurd onderwijs (PGO) en een meer gestuurde variant daarvan (GPGO) werden vergeleken in termen van leermotivatie en conceptueel begrip van de statistiek, voor studenten met meer voorkennis en studenten met minder voorkennis. In tegenstelling tot klassiek PGO zijn de leerdoelen in GPGO niet het product van groepsdiscussie onder studenten; zij zijn voorgeprogrammeerd door de docent(en) en dat voorprogrammeren is gebaseerd op een gedetailleerde ontleding van de studiestof. Deze leerdoelen kunnen dan door de GPGO groep worden gebruikt om voorkennis te activeren en om zelfstudie en daaropvolgende groepsdiscussie te structureren. De studenten die aan het experiment deelnamen werden via het lot toegewezen aan PGO dan wel GPGO. De resultaten laten zien dat, gemiddeld genomen, GPGO meer bijdraagt aan conceptueel begrip van de statistiek dan PGO zodra studenten enige voorkennis van de studiestof hebben. Bovendien lijkt GPGO een positief effect te hebben op de mate waarin studenten de waarde van het leren van statistiek inzien. Indien PGO groepen worden voorafgegaan door colleges waarin de studiestof wordt geïntroduceerd – en waarin studenten al enige voorkennis kunnen activeren – lijkt GPGO meer geschikt als instructiemethode voor het statistiekonderwijs dan klassiek PGO.

In **hoofdstuk 8** wordt een strategie voor de onderwijspraktijk gepresenteerd. Differentieel onderwijs voor studenten met weinig en veel voorkennis is niet noodzakelijk; de sleutel is afnemende sturing in instructie naarmate studenten meer (voor)kennis ontwikkelen. Ieder onderwerp dient te worden geïntroduceerd in een interactief college waarin MPM-leertaken worden gepresenteerd en/of bediscussieerd en vervolgens kunnen studenten aan (uitgewerkte voorbeelden van) MPM-leertaken in (GPGO) groepen werken. En individueel werken aan dergelijke leertaken is dan weer de volgende stap. Tijdens interactieve colleges en tijdens GPGO discussies kunnen studenten die voldoende voorkennis hebben hun kennis en begrip verbeteren door uitleg te geven aan hun medestudenten die minder voorkennis hebben. Laatstgenoemden worden zo in de gelegenheid gesteld belangrijke gaten in hun kennis en begrip weg te werken en langzaam, net als de aanvankelijk meer kennis hebbende medestudenten, meer kennis en begrip te ontwikkelen.

References

- Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, *26*, 147-179.
- Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology*, *95*, 774-783.
- Barrows, H. S. (1984). A specific, problem-based self-directed learning method designed to teach medical problem-solving skills, and enhance knowledge retention and recall. In H. G. Schmidt & M. L. de Volder (Eds.), *Tutorials in Problem-Based Learning*. Assen, The Netherlands: Van Gorcum.
- Barrows, H. S. (1996). Problem-based learning in medicine and beyond: a brief overview. In L. Wilkerson & W. H. Gijsselaers (Eds.), *New Directions for Teaching and Learning*, *68* (pp. 3-11). San Francisco: Jossey-Bass Publishers.
- Beckmann, J. F. (2010). Taming a beast of burden – on some issues with the conceptualisation of cognitive load. *Learning and Instruction*, *20*, 250-264.
- Ben-Zvi, D. & Garfield, J. B. (2004). *The challenge of developing statistical literacy, reasoning and thinking*. Dordrecht: Kluwer Academic Publishers.
- Berthold, K., & Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. *Journal of Educational Psychology*, *101*, 70-87.
- Berthold, K., Eysink, T. H. S., & Renkl, A. (2009). Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instructional Science*, *37*, 345-363.
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. P. Shimamura (Eds.), *Metacognition: Knowing about Knowing* (pp. 185-205). Cambridge, MA: MIT Press.
- Broers, N. J. (2002). Selection and use of propositional knowledge in statistical problem solving. *Learning and Instruction*, *12*, 323-344.
- Broers, N. J. (2008). Helping students to build a conceptual understanding of elementary statistics, *The American Statistician*, *62*, 1-6.
- Broers, N. J. (2009). Using propositions for the assessment of structural knowledge. *Journal of Statistics Education*, *17*, 1-19.
- Broers, N. J., & Imbos, Tj. (2005). Charting and manipulating propositions as methods to promote self-explanation in the study of statistics. *Learning and Instruction*, *15*, 517-538.
- Broers, N. J., Mur, M. C., & Bude, L. (2005). Directed self-explanation in the study of statistics. In G. Burrill & M. Camden (Eds.) *Curricular development in statistics education* (pp. 21-35). Voorburg, The Netherlands: International Statistical Institute.
- Bruinsma, M. (2003). Leidt hogere motivatie tot betere prestaties? Motivatie, informatieverwerking, en studievoortgang in het hoger onderwijs. [Does higher motivation result in higher achievement? Motivation, cognitive processing and achievement in higher education], *Pedagogische Studien*, *80*, 226-238.
- Bude, L. (2007). On the improvement of students' conceptual understanding in statistics education. PhD Dissertation, Maastricht University. Maastricht: Maastricht University Press.
- Bude, L., Imbos, Tj., Van de Wiel, M. W. J., Berger, M. P. F. (2009). The effect of directive tutor guidance in problem-based learning of statistics on students' perceptions and achievement. *Higher Education*, *57*, 23-26.
- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., Rohrer, D. (2006). Distributed practice in verbal recall tasks: a review and quantitative synthesis. *Psychological Bulletin*, *132*, 354-380.
- Chang, E., Yew, E. H. J., Schmidt, H. G. (2011). Effects of tutor-related behaviors on the process of problem-based learning. *Advances in Health Sciences Education*, *16*, 491-503.

- Chi, M. T. H. (2000). Self-explaining expository texts: the dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in Instructional Psychology* (pp. 161-238). Mahwah, NJ: Lawrence Erlbaum Associates.
- Chi, M. T. H. (2009). Active-constructive-interactive: a conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*, 73-105.
- Chi, M. T. H., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439-477.
- Cooper, G., Tindall-Ford, S., Chandler, P., & Sweller, J. (2001). Learning by imagining. *Journal of Experimental Psychology, 7*, 68-82.
- Deci, E. L., Eghrari, H., Patrick, B. C., Leone, D. (1994). Facilitating internalization: the self-determination theory perspective, *Journal of Personality, 62*, 119-142.
- DeLoach, L. J., Higgins, M. S., Caplan, A. B., & Stiff, J. L. (1998). The visual analogue scale in the immediate postoperative period: intrasubject variability and correlation with a numeric scale. *Anesthesia & Analgesia, 86*, 102-106.
- Dempster, F. N. (1988). The spacing effect: a case study in the failure to apply the results of psychological research. *American Psychologist, 43*, 627-634.
- Dolmans, D. H. J. M., De Grave, W. S., Wolfhagen, I. H. A. P., Van der Vleuten, C. P. M. (2005). Problem-based learning: future challenges for educational practice and research. *Medical Education, 39*, 732-741.
- Eysink, T. H. S., De Jong, T., Berthold, K., Kollöffel, B., Opfermann, M., & Wouters, P. (2009). Learner performance in multimedia learning arrangements: an analysis across instructional approaches. *American Educational Research Journal, 46*, 1107-1149.
- Fischer, F. (2002). Gemeinsame Wissenskonstruktion – Theoretische und methodologische Aspekte [Joint knowledge construction – theoretical and methodological aspects], *Psychologische Rundschau, 53*, 119-134.
- Fonseca, B. A., & Chi, M. T. H. (2011). Instruction based on self-explanation. In R. Mayer, & P. Alexander (Eds.), *The Handbook of Research on Learning and Instruction* (Ch. 15). Routledge Press.
- Gal, I., & Ginsburg, L. (1994). The role of beliefs and attitudes in learning statistics: towards an assessment framework, *Journal of Statistics Education, 2* [Online].
- Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*, 22-38.
- Hausmann, R. G., Van de Sande, B., & Van Lehn, K. (2008). Are self-explaining and coached problem solving more effective when done by pairs of students than alone? In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, (pp. 2369-2374). New York, NY: Erlbaum.
- Huberty, C. J., Dresden, J., & Bak, B. (1993). Relations among dimensions of statistical knowledge. *Educational and Psychological Measurement, 53*, 523-532.
- Hulsizer, M. R., & Woolf, L. M. (2009). *A guide to teaching statistics: innovations and best practices*. Oxford: Wiley-Blackwell.
- Johnson, D. W., Johnson, R. T., & Smith, K. (2007). The state of cooperative learning in postsecondary and professional settings. *Educational Psychology Review, 19*, 15-29.
- Kalyuga, S. (2005). Prior knowledge principle. In R. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 325–337). New York: Cambridge University Press.
- Kalyuga, S. (2006). *Instructing and testing advanced learners: A cognitive load approach*. New York: Nova Science.
- Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*, 509-539.
- Kalyuga, S. (2009). Knowledge elaboration: a cognitive load perspective. *Learning and Instruction, 19*, 402-410.
- Kalyuga, S., & Hanham, J. (2010). Instructing in generalized knowledge structures to develop flexible problem solving skills. *Computers in Human Behavior*, doi:10.1016/j.chb.2010.05.024.

- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect, *Educational Psychologist*, *38*, 23-31.
- Kalyuga, S., Chandler, P., & Sweller, J. (2001a). Learner experience and efficiency of instructional guidance. *Educational Psychology*, *21*, 5-23.
- Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001b). When problem solving is superior to studying worked examples. *Journal of Educational Psychology*, *93*, 579-588.
- Kaplan, J. J. (2006). *Factors in statistics learning: developing a dispositional attribution model to describe differences in the development of statistical proficiency*. PhD Dissertation.
- Karpicke, J. D., & Roediger, H. L., 3rd. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966-968.
- King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Educational Research Journal*, *29*, 303-323.
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based learning. *Educational Psychologist*, *41*, 75-86.
- Knipfer, K., Mayr, E., Zahn, C., Schwan, S., Hesse, F. W. (2009). Computer support for knowledge communication in science exhibitions: novel perspectives from research on collaborative learning. *Educational Research Review*, *4*, 196-209.
- Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*, 219-224.
- Kramarski, B., & Dudai, V. (2009). Group-metacognitive support for online inquiry in mathematics with differential self-questioning. *Journal of Educational Computing Research*, *40*, 377-404.
- Krol, K., Janssen, J., Veenman, S., & Van der Linden, J. (2004). Effects of a cooperative learning program on the elaborations of students in dyads. *Educational Research and Evaluation*, *10*, 205-237.
- Kuhl, J. (2000). A functional-design approach in motivation and self-regulation: the dynamics of personality systems interactions. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 111-169). San Diego, CA: Academic Press.
- Larsen, D. P., Butler, A. C., & Roediger, H. L., 3rd. (2008). Test-enhanced learning in medical education. *Medical Education*, *42*, 959-966.
- Lee, G. H., Lin, Y. H., Tsou, K. I., Shiau, S. J., & Lin, C. S. (2009). When a problem-based learning tutor decides to intervene. *Academic Medicine*, *84*, 1406-1410.
- Leppink, J. (2010). Adjusting cognitive load to the student's level of expertise for increasing motivation to learn, *Proceedings of the Eighth International Conference on Teaching Statistics, Ljubljana, Slovenia*.
- Leppink, J. (2011). Zelfverklaring door middel van argumentatie: de invloed van voorkennis [Self-explanation by means of argumentation: the influence of prior knowledge], *Onderwijs Research Dagen 2011, Maastricht, The Netherlands*.
- Leppink, J., Broers, N. J., Imbos, Tj., Van der Vleuten, C. P. M., & Berger, M. P. F. (2011a). Exploring task- and student-related factors in the method of propositional manipulation (MPM). *Journal of Statistics Education*, *19*, 1-23.
- Leppink, J., Broers, N. J., Imbos, Tj., Van der Vleuten, C. P. M., & Berger, M. P. F. (2011b). Self-explanation in the domain of statistics: an expertise reversal effect, *Higher Education*, DOI 10.1007/s10734-011-9476-1.
- Leppink, J., Broers, N. J., Imbos, Tj., Van der Vleuten, C. P. M., & Berger, M. P. F. (2012). Prior knowledge moderates instructional effects on conceptual understanding of statistics, *Educational Research and Evaluation*, *18*, 37-51.
- Levesque, C., Zuehlke, A. N., Stanek, L. R., & Ryan, R. M. (2004). Autonomy and competence in German and American university students: a comparative study based on self-determination theory. *Journal of Educational Psychology*, *96*, 68-84.
- Lovett, M. C. (1992). Learning by problem solving versus by examples: the benefits of generating and receiving information. In *Proceedings of the 14th annual conference of the Cognitive Science Society* (pp.

- 956e961). Hillsdale, NJ: Erlbaum. Miyake, A., & Shah, P. (1999). Models of working memory: mechanisms of active maintenance and executive control. Cambridge, England: Cambridge University Press.
- Machiels-Bongaerts, M., Schmidt, H. G., & Boshuizen, H. P. A. (1995). The effect of prior knowledge activation on text recall: an investigation of two conflicting hypotheses. *British Journal of Educational Psychology*, *65*, 409-423.
- Markland, D., & Hardy, L. (1997). On the factorial and construct validity of the intrinsic motivation inventory: conceptual and operational concerns. *Research Quarterly for Exercise and Sports*, *68*, 20-32.
- Marshall, S. (1995). *Schemas in problem solving*. Cambridge: Cambridge University Press.
- Martens, R. L., Gulikers, J., & Bastiaens, T. (2004). The impact of intrinsic motivation on e-learning in authentic computer tasks. *Journal of Computer Assisted Learning*, *20*, 368-376.
- Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). New York: Freeman.
- McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: a confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, *60*, 48-58.
- Miyake, A., & Shah, P. (1999). *Models of working memory: mechanisms of active maintenance and executive control*. Cambridge, England: Cambridge University Press.
- Moore, D. S., McCabe, G. P., & Craig, B. (2009). *Introduction to the practice of statistics* (6th ed.). New York: Freeman.
- Moreno, R. (2009). Constructing knowledge with an agent-based instructional program: a comparison of cooperative and individual meaning making. *Learning and Instruction*, *19*, 433-444.
- Myles, P. S., Troedel, S., Boquest, M., & Reeves, M. (1999). The pain visual analog scale: is it linear or nonlinear? *Anesthesia & Analgesia*, *89*, 1517-1520.
- Norman, G. R., & Schmidt, H. G. (2000). Effectiveness of problem-based learning curricula: theory, practice and paper darts. *Medical Education*, *34*, 721-728.
- Novak, J. D. (2002). Meaningful learning: the essential factor for conceptual change in limited or inappropriate propositional hierarchies leading to empowerment of learners. *Learning*, 548-571.
- Nussbaum, E. M. (2008). Collaborative discourse, argumentation, and learning: preface and literature review. *Contemporary Educational Psychology*, *33*, 345-359.
- O'Donnell, A. M. (2006). The role of peers and group learning. In P. H. Winne & P. A. Alexander (Eds.), *Handbook of Educational Psychology* (pp. 781-802). Mahwah, NJ: Erlbaum.
- Paas, F. (1992). Training strategies for attaining transfer of problem-solving skills in statistics: a cognitive load approach. *Journal of Educational Psychology*, *84*, 429-434.
- Paas, F., & Van Gog, T. (2006). Optimising worked example instruction: different ways to increase germane cognitive load. *Learning and Instruction*, *16*, 87-91.
- Paas, F., & 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.
- Peeck, J. (1982). Effects of mobilization of prior knowledge on free recall, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*, 608-612.
- Pichert, J. W., & Anderson, R. C. (1977). Taking different perspectives on a story. *Journal of Educational Psychology*, *69*, 309-315.
- Roscoe, R. D., & Chi, M. T. H. (2008). Tutor learning: the role of explaining and responding to questions. *Instructional Science*, *36*, 321-350.
- Ryan, R. M. (1982). Control and information in the intrapersonal sphere: an extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, *43*, 450-461.
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*, 68-78.

- Ryan, R. M., Koestner, R., & Deci, E. L. (1991). Varied forms of persistence: when free-choice behavior is not intrinsically motivated. *Motivation and Emotion, 15*, 185-205.
- Ryan, R. M., Mims, V., & Koestner, R. (1983). Relation of reward contingency and interpersonal context to intrinsic motivation: a review and test using cognitive evaluation theory. *Journal of Personality and Social Psychology, 45*, 736-750.
- Schmidt, H. G., Van der Molen, H. T., Te Winkel, W. W. R., Wijnen, W. H. F. W. (2009). Constructivist, problem-based learning does work: a meta-analysis of curricular comparisons involving a single medical school. *Educational Psychologist, 44*, 227-249.
- Schnotz, W., & Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review, 19*, 469-508.
- Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: a meta-analysis. *Review of Educational Research, 69*, 21-51.
- Sweller, J., & Van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-296.
- Tsigilis, N., & Theodosiou, A. (2003). Temporal stability of the intrinsic motivation inventory, *Perceptual and Motor Skills, 97*, 271-280.
- Van Blankenstein, F. M., Dolmans, D. H. J. M., Van der Vleuten, C. P. M., & Schmidt, H. G. (2009). Which cognitive processes support learning during small-group discussion? The role of providing explanations and listening to others. *Instructional Science, 39*, 189-204.
- Van Buuren, J. A. (2008). *Van vakgericht naar competentiegericht statistiekonderwijs: een interventiestudie in een opleiding psychologie* [From subject-oriented to competence-based statistics education: an intervention study in a school of psychology] PhD Dissertation, Open University of The Netherlands. Voerendaal: Schrijen Lippertz Huntjens.
- Van Merriënboer, J. J. G., Schuurman, J. G., De Croock, M. B. M., & Paas, F. (2002). Redirecting learners' attention during training: effects on cognitive load, transfer test performance, and training efficiency. *Learning and Instruction, 12*, 11-37.
- Van Merriënboer, J., & Sweller, J. (2005). Cognitive load theory and complex learning: recent developments and future directions. *Educational Psychology Review, 17*, 147-177.
- Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Education Research, 13*, 21-39.
- Wetzels, S. A. J., Kester, L., & Van Merriënboer, J. J. G. (2011). Adapting prior knowledge activation: mobilization, perspective taking, and learners' prior knowledge. *Computers in Human Behavior, 27*, 16-21.
- Wetzels, S. A. J. (2009). *Individualised strategies for prior knowledge activation* [dissertation], Netherlands: Interuniversity Centre for Education Research (ICO).
- Willoughby, T., Waller, T. G., Wood, E., & MacKinnon, G. E. (1993). The effect of prior knowledge on an immediate and delayed associative learning task following elaborative interrogation, *Contemporary Educational Psychology, 18*, 36-46.
- Woloshyn, V. E., Pressley, M., & Schneider, W. (1992). Elaborative interrogation and prior knowledge effects on learning of facts, *Journal of Educational Psychology, 84*, 115-124.

Appendix

More learning materials (not displayed in the previous chapters)

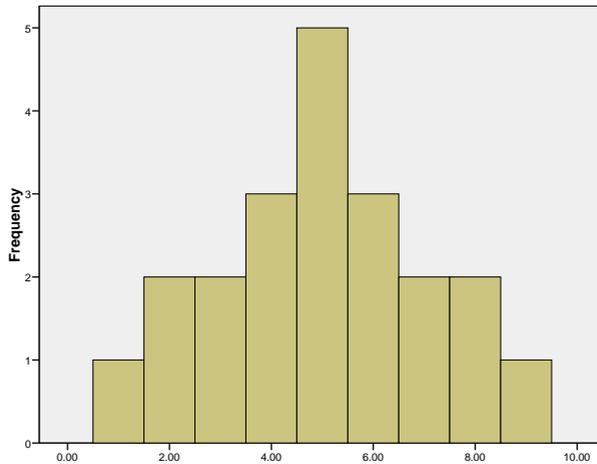
1 See the cross table below for the association between gender and hair color in a population, in which only three colors of hair exist: blond, brown, and black. In this population, for both men and women the univariate distribution for hair color is: 30% blond, 50% brown and 20% black.

	Female	Male
Blond	45	60
Brown	75	100
Black	30	40
Total	150	200

Hypothesis: From the information it can be concluded that in this population, there is no association between gender and hair color.

- [1] What is the *marginal distribution* of a variable?
- [2] What is the *conditional distribution* of a variable?
- [3] When can we say there is an *association* between two categorical variables?

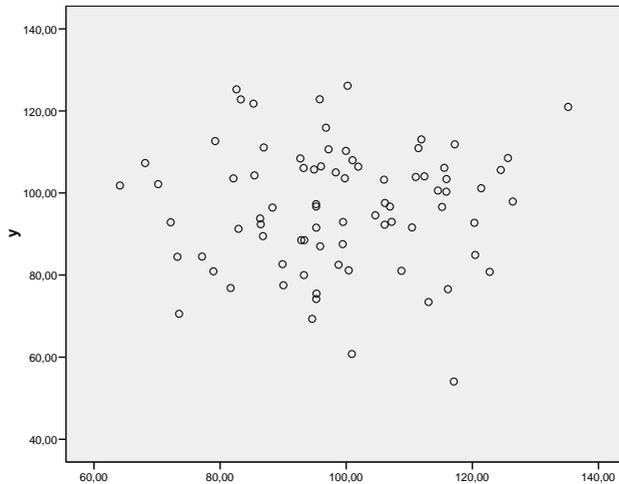
2 The histogram below shows the distribution of a variable x . The arithmetic mean of this distribution equals 5 and its standard deviation is 2.12, hence the length of the interval '*arithmetic mean plus or minus one time the standard deviation*' equals 4.24. The first quartile (Q_1) equals 3.5, the third quartile (Q_3) equals 6.5.



Hypothesis: For this distribution of variable x , the interval '*arithmetic mean plus or minus one time the standard deviation*' contains more than 50% of the values.

- [1] How is the *interquartile range (IQR)* computed from the values of Q_1 and Q_3 ?
- [2] What range of values is represented by the *IQR*?
- [3] What is the definition of the *median*?
- [4] What can be said about median and arithmetic mean in the case of a *perfectly symmetrical distribution*?

3 The scatterplot below is about the association between two variables x and y . For this distribution, the correlation coefficient r_{xy} is equal to .03.



Hypothesis: In this case, the correlation coefficient r_{xy} does not give a good summary of the association between x and y .

- [1] What is expressed by the correlation coefficient r_{xy} ?
- [2] What values can the correlation coefficient r_{xy} have?
- [3] What is an *outlier* in a scatterplot?
- [4] How can an outlier influence the value of the correlation coefficient r_{xy} ?
- [5] If the association between two variables x and y is *non-linear*, what can be said about the value of the correlation coefficient r_{xy} ?

4 Imagine we test a one-sided hypothesis, our test statistic has a certain value, and the accompanying P -value equals .07. Our significance level is .05.

Hypothesis: The P -value of .07 does not give rise to a Type I error, but does give rise to a Type II error.

- [1] What is a P -value?
- [2] What is meant by *significance level*?
- [3] What is a *Type I error*?
- [4] What is a *Type II error*?

5 Hypothesis: If the value of our test statistic exceeds the critical value, the P -value is smaller than the significance level.

- [1] What is a *test statistic*?
- [2] What is a *sampling distribution*?
- [3] What is a P -value?
- [4] What is meant by *significance level*?
- [5] What is meant by *critical value* in the context of hypothesis testing?

Acknowledgements

In writing a thesis, you are never alone. I thank my supervisors Martijn Berger and Cees van der Vleuten, my official co-supervisor Nick Broers, and 'unofficial second co-supervisor' Tjaart Imbos. First of all, thank you for the trust each of you has put in me. Martijn, your methodological and statistical considerations have really improved the quality of the studies presented in this thesis. Besides, your 'optimal design approach' with regard to very long texts has helped me shorten some of the chapters considerably without losing content. Further, thank you for the conversations on statistics courses and my future. Finally, thank you all for the opportunities you have given me in the Department of Methodology and Statistics. Cees, what Martijn has contributed from the methodological and statistical point of view, you have contributed from the educational theories point of view. Your contributions to the theoretical framework of this thesis have been great. We had many discussions about educational theory, about the implications of the studies presented in this thesis for further research, and you strongly encouraged me to continue educational research after this PhD project. Thank you for all the opportunities you have given me in the Department of Educational Development and Research, thank you for our conversations about my future, and I hope the AMEE research guide and other projects are going to be a success. Nick, you are the architect of the main instructional method in this thesis, so this project would not have existed without you. Your achievements as a teacher deserve respect and so does the development of MPM. Even if the PhD candidate in the project brings in new ideas that shape the studies carried out in that project, in the end the foundation is still the same. Thank you for all the ideas you have shared, and thank you for the trust you have put in me even in the application rounds for this project. Tjaart, thank you for the trust you have put in me since the application rounds for this project. As a nationally and internationally recognized pioneer in statistics education research, your contributions in this project have improved each the studies presented in this thesis from both methodological and theoretical perspective. You have a passion for teaching statistics and you recognize the importance of statistics education research more than almost anyone in the statistics knowledge domain.

I thank my colleagues from the Department of Methodology and Statistics, especially Frans Tan, Gerard van Breukelen, Haftom Temesgen Abebe, Math Candel, Mickey Chenault, Monique Reusken, Mary Duryan, and Marga Doyle. Frans and Gerard, thank you for all the insights with regard to multilevel modeling. Frans, thank you also for your statistical advice with regard to Chapter 5 of this thesis. Haftom and Math, thank you for all the discussions we had. Some of them have given me inspiration for further research. Math, also thank you for the trust you have put in me in the application rounds for this project. Mickey, thank you for all the discussions with regard to statistics and with regard to education. You have lots of experience and it is always nice to discuss with you. Mary and Monique, thank you for all the conversations and discussions, some of which have contributed to this thesis as well. It was nice to share the office. Marga, thank you for all the administrative support throughout the project and for your suggestions with regard to the formulation of some of the propositions (in Dutch: 'stellingen behorende bij dit proefschrift').

I thank my colleagues from the Department of Educational Development and Research, especially Floris van Blankenstein, Renee Stalmeijer, Ingrid Spanjers, Jeroen van Merriënboer,

Arno Muijtjens, Diana Dolmans, Lilian Swaen, Nicky Verleng, and Stefan Groenveld. Floris, Renee, Ingrid, and Jeroen, each of you has contributed to this thesis even if you may not be aware of it. Floris, thank you for the conversations about randomized experiments in educational research. Renee, thank you for the discussions on qualitative and mixed methods research. Ingrid, thank you for the discussions, references, and materials on cognitive load. Jeroen, thank you for the discussions on cognitive load, and for the opportunities you have given me in the Department of Educational Development and Research. I am very happy to be involved in the methodology and statistics part in a variety of research projects, and the cooperation with Arno and Renee is great. Arno, thank you for all the discussions we had on methodology and statistics in education research. As in any area of social science research, communication about methodology and statistics in education research requires experience. You have that experience, and I have learned a lot from you. I hope that our AMEE research guide and other projects are going to be a success. Diana, thank you for the discussions during the PhD reading club sessions, and thank you for giving me the opportunity to present some of my research in the ICO Master class. Lilian, Nicky, and Stefan, thank you for all the administrative support throughout the project.

I thank the members of the thesis evaluation committee, Diana Dolmans, Arthur Bakker, Wim Gijsselaers, Fred Paas, and Dirk Tempelaar for taking their time to evaluate this thesis. Also, a special thank you to Dirk Tempelaar, Wim Gijsselaers, and Arthur Bakker for their very useful comments.

I thank Latifa Abidi, Patrick Schwarz, and Jana Leppink for their time, contribution to data analysis, ideas, and good discussions. Jana, you are the best wife I could have ever imagined. You give me warmth and color. There are not so many things in life you can be certain about, but I am certain about you; we are a team no matter what comes.

I thank my parents Leon and Marion Leppink for giving me this wonderful life full of events and with virtually limitless possibilities, for investing in my education, and for all the encouragements to keep learning. Petru and Valentina Gorodenco, I am lucky to call you parents as well and your place feels like a second home. Marina Gorodenco, you feel like a sister. We have had many long discussions about lots of things in life, share quite some interests, and doing our PhD at the same time, it felt like we were in the same boat.

About the author

Jimmie Leppink was born in Heerlen, the Netherlands, on April 28, 1983. He obtained a BSc in Cognitive Psychology in 2005, an MSc in Psychology and Law in 2006 (Cum Laude), and an LLM in Forensics, Criminology, and Law in 2008, at Maastricht University, the Netherlands. In September 2008, he started his PhD in Statistics Education at the Department of Methodology and Statistics and the Department of Educational Development and Research, at Maastricht University. His research focused on instructional methods for statistics in the social sciences and health sciences. He completed extracurricular courses in statistics, education research, qualitative research methodology, entrepreneurship, and legal psychology. Keen on languages, he completed courses in Spanish, Italian, Portuguese, Turkish, Russian, Romanian, and French, while using English, Dutch, and German in professional activities. He has a passion for teaching in methodology and statistics; he won Maastricht University's education award of best tutor in the health sciences for the year 2011. He is currently completing an MSc program in Statistics (Quantitative Analysis in the Social Sciences) at Catholic University of Leuven, Belgium, and he is doing research on a new instrument for the measurement of cognitive load in the statistics knowledge domain.

SHE dissertation series

In the SHE Dissertation Series dissertations are published of PhD candidates from the School of Health Professions Education (SHE) who defended their PhD thesis at Maastricht University. The most recent ones are listed below. For more information go to: www.maastrichtuniversity.nl/she.

Claramita, M. (30-03-2012) Doctor-patient communication in a culturally hierarchical context of Southeast Asia: A partnership approach

Kleijnen, J. (21-03-2012) Internal quality management and organisational values in higher education

Persoon, M.C. (19-01-2012) Learning in Urology; The influence of simulators and human factors

Pawlikowska, T.R.B. (21-12-2011) Patient Enablement; A Living Dialogue

Sok Ying Liaw, (14-12-2011) Rescuing A Patient In Deteriorating Situations (RAPIDS): A programmatic approach in developing and evaluating a simulation-based educational program

Singaram, V.S. (7-12-2011) Exploring the Impact of Diversity Factors on Problem-Based Collaborative Learning

Balslev, T. (24-11-2011) Learning to diagnose using patient video cases in paediatrics: Perceptive and cognitive processes

Widyandana, D. (19-10-2011) Integrating Pre-clinical skills training in skills laboratory and primary health care centers to prepare medical students for their clerkships

Durning, S.J. (09-09-2011) Exploring the Influence of Contextual Factors of the Clinical Encounter on Clinical Reasoning Success (Unraveling context specificity)

Govaerts, M.J.B. (08-09-2011) Climbing the Pyramid; Towards Understanding Performance Assessment

Stalmeijer, R. E. (07-07-2011) Evaluating Clinical Teaching through Cognitive Apprenticeship.

Malling, B.V.G. (01-07-2011) Managing word-based postgraduate medical education in clinical departments

Veldhuijzen, J.W. (17-06-2011) Challenging the patient-centred paradigm: designing feasible guidelines for doctor patient communication.

Van Blankenstein, F. (18-05-2011) Elaboration during problem-based, small group discussion: A new approach to study collaborative learning.

Van Mook, W. (13-05-2011) Teaching and assessment of professional behavior: Rhetoric and reality.

De Leng, B. (8-12-2009). Wired for learning. How computers can support interaction in small group learning in higher education.

Maiorova, T. (29-05-2009). The role of gender in medical specialty choice and general practice preferences.

Bokken, L. (04-03-2009). Innovative use of simulated patients for educational purposes.

Wagenaar, A. (18-09-2008). Learning in internships. What and how students learn from experience.

Driessen, E. (25-06-2008). Educating the self-critical doctor. Using portfolio to stimulate and assess medical students' reflection.

Derkx, H. (18-06-2008). For your ears only. Quality of telephone triage at out-of-hours centres in the Netherlands.

Niessen, Th. (30-11-2007). Emerging epistemologies: making sense of teaching practice.

Budé, L. (05-10-2007). On the improvement of students' conceptual understanding in statistics education.

Niemantsverdriet, S. (26-07-2007). Learning from international internships: A reconstruction in the medical domain.

Marambe, K. (20-06-2007). Patterns of student learning in medical education – A Sri Lankan study in traditional curriculum.

Pleijers, A. (19-01-2007). Tutorial group discussion in problem-based learning.

Sargeant, J. (21-09-2006). Multi-source feedback for physician learning and change.

Dornan, T. (12-06-2006). Experience-based learning.

Wass, V. (12-05-2006). The assessment of clinical competence in high stakes examinations.

Prince, K. (21-04-2006). Problem-based learning as a preparation for professional practice.