# Self-explanation in the domain of statistics: an expertise reversal effect

**Jimmie Leppink · Nick J. Broers · Tjaart Imbos ·
Cees P. M. van der Vleuten · Martijn P. F. Berger**

**Abstract** This 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. The instructional methods were (1) a reading-only control condition, (2) answering open-ended questions, (3) answering open-ended questions and formulating arguments, and (4) 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 learn most from studying worked-out examples, whereas 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.

**Keywords** Cognitive load · Propositional knowledge · Conceptual understanding · Expertise reversal effect

## 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

J. Leppink (✉) · N. J. Broers · T. Imbos · M. P. F. Berger
Department of Methodology and Statistics, Maastricht University, P. O. Box 616,
6200 MD Maastricht, The Netherlands
e-mail: j.leppink@maastrichtuniversity.nl

C. P. M. van der Vleuten
Department of Educational Development and Research, Maastricht University, Maastricht,
The Netherlands

knowledge (i.e., conceptual understanding, Broers 2009). Knowledge elaboration can help students develop the latter ability (Kalyuga 2009).

Knowledge elaboration

Knowledge elaboration is using prior knowledge to structure and integrate new information. Self-explanation (Atkinson et al. 2003; Berthold and 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 Merrienboer and Sweller 2005). Working memory is limited in capacity (Miller 1956) as well as in duration (Miyake and Shah 1999a, b). 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 Merrienboer and Sweller 2005). Most people can retain $7 \pm 2$ 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 and 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.

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 and 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 and 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 and Van Merrienboer 1994; Van Merrienboer 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 and 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. 2003, 2001a).

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 Merrienboer and 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 Merrienboer and 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.

## 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.

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.

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 and McCabe (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 min test consisting of a part measuring propositional knowledge and another part measuring conceptual understanding.

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 min. 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 min 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 1.

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 2.

Each learning task comprised a true/false statement and three of the nine open-ended questions (as presented in Box1), 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 form the open-ended questions plus arguments condition in the form of worked-out examples. For a worked-out example of an argument see Box 3.

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 min studying session, students did a 50 min 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.

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

**Box 1** Example of open-ended questions

[1] What is meant by the sampling distribution of a statistic?

[2] What is meant by the expected value of the statistic?

[3] What is meant by variation in the values of the statistic?

---

**Box 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.

*Statement* The sample mean equals 10.

Questions to be processed in the argument:

[1] What is meant by the sampling distribution of a statistic?

[2] What is meant by the expected value of the statistic?

[3] What is meant by variation in the values of the statistic?

---

**Box 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. A statistic like the sample mean varies over these possible samples and thus has variation [*answer to question 3 in* Box 2]. The sampling distribution of a statistic is the probability distribution of the values of the statistic over all possible samples of the same sample size $N$ from the same population [*answer to question 1 in* Box 2]. Thus, the sampling distribution is about variation in the statistic [*answers to questions 1 and 3 related*]. The expected value of the statistic is the average value the statistic over all possible samples of the same sample size $N$ from the same population [*answer to question 2 in* Box 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 statement is false.

---

**Box 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

---

load was measured twice, namely at the end of the studying session and at the end of the test.

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 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 $\kappa = 0.87$), the correlation between the sum scores of the two raters was also high ($r = 0.97$) 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 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.

## Results

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

### 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 of cognitive load when studying and cognitive load during the exam are presented in Table 1.

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, $F(3, 97) = 0.175$, $p > .90$ ($\eta^2 = 0.005$) as well as a non-significant two-way interaction between cognitive load and prior knowledge group, $F(1, 97) = 1.087$, $p > .30$ ($\eta^2 = 0.011$). However, the interaction between cognitive load and instructional method was statistically significant, $F(3, 101) = 3.690$, $p < .05$ ($\eta^2 = 0.099$). 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, $F(3, 97) = 0.684$, $p > .55$ ($\eta^2 = 0.021$). However, the main effect of instructional method was significant, $F(3, 101) = 2.991$, $p < .05$ ($\eta^2 = 0.082$). 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, $F(1, 103) = 0.073$,

**Box 5** Example of how the open-ended questions were scored

*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

**Table 1** Means (and standard deviations) and sample sizes ($n$) of cognitive load when studying

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

$p > .75$ ($\eta^2 = 0.001$). 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, $F(3, 97) = 0.348$, $p > .75$ ($\eta^2 = 0.011$). The same holds for the main effect of instructional method, $F(3, 101) = 0.837$, $p > .45$ ($\eta^2 = 0.024$), as well as for the main effect of prior knowledge, $F(1, 103) = 0.707$, $p > .40$ ($\eta^2 = 0.007$).

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.

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 2.

The interaction effect was very small and not statistically significant, $F(3, 97) = 0.389$, $p > .75$ ($\eta^2 = 0.012$). The same holds for the main effect of instructional method, $F(3,$

**Table 2** Means (and standard deviations) 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) | $n = 12$ | 5.33 (2.23) | $n = 12$ |
| Questions | 5.29 (2.64) | $n = 14$ | 6.15 (2.15) | $n = 13$ |
| Questions and arguments | 4.57 (1.83) | $n = 14$ | 6.14 (1.92) | $n = 14$ |
| Worked-out examples | 5.23 (2.74) | $n = 13$ | 5.46 (2.15) | $n = 13$ |
| Total | 4.87 (2.42) | $n = 53$ | 5.79 (2.08) | $n = 52$ |

101) $= 0.609$, $p > .60$ ($\eta^2 = 0.018$). The main effect of prior knowledge was statistically significant, $F(1, 103) = 4.366$, $p < .05$ ($\eta^2 = 0.041$). 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 op 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 $d = 0.41$).

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, $F(3, 97) = 2.762$, $p < .05$ ($\eta^2 = 0.079$). Means and standard deviations of conceptual understanding are presented in Table 3.

Given the interaction pattern, main effects are difficult to interpret. Although two-way ANOVA yields a marginally significant main effect for prior knowledge, $F(1, 97) = 3.32$, $p < .10$ ($\eta^2 = 0.033$) and also a small to medium size (although non-significant) main effect for instructional method, $F(3, 97) = 1.625$, $p < .20$ ($\eta^2 = 0.048$), 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, $F(3, 48) = 2.219$, $p < .10$ ($\eta^2 = 0.122$), and for low prior knowledge students, $F(3, 49) = 2.097$, $p < .15$ ($\eta^2 = 0.114$).

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.

Table 3 Means (and standard deviations) 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) | $n = 12$ | 5.42 (1.56) | $n = 12$ |
| Questions | 5.14 (1.70) | $n = 14$ | 5.00 (2.08) | $n = 13$ |
| Questions and arguments | 4.64 (1.08) | $n = 14$ | 6.79 (1.31) | $n = 14$ |
| Worked-out examples | 6.00 (1.78) | $n = 13$ | 5.85 (2.41) | $n = 13$ |

## 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.

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 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 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 and Van Merrienboer 1993; Van Gog and 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.

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 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 min. 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.

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 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 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 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.

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 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 2 for an example) imposes germane load on students, while reducing overall load (see Table 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 3): students are confronted with worked-out examples
- *Semi-guided* or *self-guided problem-solving* (e.g., Box 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)
- *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. Aleven 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.

# 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.

Beckmann, J. F. (2010). Taming a beast of burden–on some issues with the conceptualisation of cognitive load. *Learning and instruction, 20*, 250–264.

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.

Broers, N. J. (2002). *Learning statistics by manipulating propositions. Proceedings of the Sixth International Conference on Teaching Statistics.* Capetown, South Africa.

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.

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: Applied, 7*, 68–82.

Eysink, T. H. S., De Jong, T., Berthold, K., Kolloffel, 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.

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

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* (2010), doi:10.101G/j.cnb.2010.05.024.

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.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*, 23–32.

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.

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.

Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). New York: Freeman.

Miller, G. A. (1956). The magic number seven plus or minus two: Some limits in our capacity to process information. *Psychological Review, 63*, 81–97.

Miyake, A., & Shah, P. (1999a). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge, England: Cambridge University Press.

Miyake, A., & Shah, P. (1999b). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge, England: Cambridge University Press.

Moore, D. S., & McCabe, G. P. (2009). *Introduction to the practice of statistics* (6th ed.). New York: Freeman.

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 Merrienboer, 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 Merrienboer, 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.

Sweller, J., Van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296.

Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43*, 16–26.

Van Merrienboer, J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*, 147–177.

Van Merrienboer, 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.