Effects of pairs of problems and examples on task performance and different types of cognitive load

Learning and Instruction 30 (2014) 32–42

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Effects of pairs of problems and examples on task performance and different types of cognitive load

Jimmie Leppink [a,*], Fred Paas [b,c], Tamara van Gog [b], Cees P.M. van der Vleuten [a], Jeroen J.G. van Merriënboer [a]

[a] Department of Educational Development and Research, Maastricht University, The Netherlands
[b] Institute of Psychology, Erasmus University Rotterdam, The Netherlands
[c] Interdisciplinary Educational Research Institute, University of Wollongong, Australia

ARTICLE INFO

Article history:
Received 30 May 2013
Received in revised form
2 December 2013
Accepted 3 December 2013

Keywords:
Cognitive load
Example–example pairs
Example–problem pairs
Problem–example pairs
Problem–problem pairs

ABSTRACT

In two studies, we investigated whether a recently developed psychometric instrument can differentiate intrinsic, extraneous, and germane cognitive load. Study I revealed a similar three-factor solution for language learning ($n = 108$) and a statistics lecture ($n = 174$), and statistics exam scores correlated negatively with the factors assumed to represent intrinsic and extraneous cognitive load during the lecture. In Study II, university freshmen who studied applications of Bayes' theorem in example–example ($n = 18$) or example–problem ($n = 18$) condition demonstrated better posttest performance than their peers who studied the applications in problem–example ($n = 18$) or problem–problem ($n = 20$) condition, and a slightly modified version of the aforementioned psychometric instrument could help researchers to differentiate intrinsic and extraneous cognitive load. The findings provide support for a recent reconceptualization of germane cognitive load as referring to the actual working memory resources devoted to dealing with intrinsic cognitive load.

1. Introduction

The central tenet of cognitive load theory (Sweller, 2010; Sweller, Ayres, & Kalyuga, 2011; Sweller, Van Merriënboer, & Paas, 1998; Van Merriënboer & Sweller, 2005, 2010) is that human cognitive architecture – and especially the limitations of working memory – should be taken into account when designing instruction. Working memory has a limited capacity of seven plus or minus two elements (or chunks) of information when merely holding information (Miller, 1956) and even fewer (circa four) when processing information (Cowan, 2001). Working memory load (or cognitive load) is therefore determined by the number of information elements that need to be processed simultaneously within a certain amount of time (Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007). Originally, cognitive load theory distinguished between two sources of cognitive load, namely intrinsic and extraneous cognitive load (Sweller, 2010; Sweller et al., 2011, 1998).

Intrinsic cognitive load is determined by the intrinsic nature of the information to be learned, more specifically, by the number of interacting information elements that the learning task or the learning material comprises (Sweller, 1994; Sweller et al., 2011). Novices, who have little if any prior knowledge of the task or material, have to process (i.e., select, organize, and integrate) those interacting elements in order to learn the task or material. As learning progresses (i.e., expertise increases), information elements become incorporated (or chunked) into cognitive schemata stored in long-term memory, which can be handled as one single element in working memory. Therefore, the intrinsic cognitive load that is imposed by a learning task or learning materials is much higher for novices than for more advanced students.

Extraneous cognitive load arises from suboptimal instructional methods that require the learner to engage in cognitive processes that do not contribute directly to the construction of cognitive schemata (e.g., having to mentally integrate spatially or temporally separated but mutually referring information sources) and are as such unnecessary and extraneous to the learning goals (Sweller & Chandler, 1994; Sweller, Chandler, Tierney, & Cooper, 1990). Such processes can hamper learning if intrinsic cognitive load is high or

* Corresponding author. Department of Educational Development and Research, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Tel.: +31 433885709.
E-mail address: jimmie.leppink@maastrichtuniversity.nl (J. Leppink).

0959-4752/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.learninstruc.2013.12.001

lead to suboptimal learning under conditions in which intrinsic cognitive load is low. That is, even though extraneous cognitive load can be managed without hampering learning under such conditions, a replacement of the extraneous by cognitive load that is directly relevant for learning (i.e., *germane cognitive load*; Sweller et al., 1998) would have resulted in better learning outcomes.

The concept of germane cognitive load was added to the cognitive load framework later on (Sweller et al., 1998). This type of load arises from relating relevant information from long-term memory or context to the new information elements (Sweller, 2010; Sweller et al., 2011) and as such pertains to the working memory resources allocated to dealing with intrinsic cognitive load (Kalyuga, 2011; Sweller, 2010). In fact, the term 'germane cognitive load' has been used in the traditional conceptualization of cognitive load theory (Sweller et al., 1998), while the term 'germane resources' (i.e., working memory resources allocated to dealing with intrinsic cognitive load) has been used in the recent version of the theory and is thus related to intrinsic cognitive load (Kalyuga, 2011; Sweller, 2010; Sweller et al., 2011).

Cognitive load theory states that intrinsic cognitive load should be optimized in instructional design by selecting materials that match the learner's prior knowledge or proficiency, while extraneous cognitive load should be minimized, and learners should be challenged to engage in processes that evoke germane cognitive load (in the old conceptualization of cognitive load theory) or the use of (in the new conceptualization of the theory) germane resources (e.g., variability in practice, elaboration, or self-explanation) and contribute directly to the construction of cognitive schemata (Sweller et al., 1998; Van Merriënboer & Sweller, 2005, 2010). To avoid confusion due to using both terms interchangeably thereby referring to two different conceptualizations of the theory, in the remainder of this paper we use the term 'germane cognitive load' as referring to the use of germane resources, as suggested by Kalyuga (2011), Sweller (2010), and Sweller et al. (2011).

1.1. Instructional guidance and cognitive load

The extent to which instructional features contribute to intrinsic or extraneous cognitive load may depend on the individual learner. For instance, novice learners, for whom information imposes high intrinsic cognitive load, may learn better from an instructional format that reduces extraneous cognitive load, such as worked examples (i.e., fully worked-out problem solutions; Cooper & Sweller, 1987; Paas, 1992; Paas & Van Merriënboer, 1994a; Sweller & Cooper, 1985; Van Gog, Paas, & Van Merriënboer, 2006) or from completing partially worked-out solutions (i.e., completion problems; Paas, 1992; Van Merriënboer, 1990) than from autonomous problem solving. Problem solving imposes high extraneous cognitive load for novice learners, because their lack of prior knowledge of how to solve that type of problem forces them to resort to weak problem-solving strategies. Because (part of) the solution is worked out in worked examples and completion problems, the extraneous cognitive load imposed by the use of weak problem-solving strategies is prevented, and learners can allocate more of their working memory resources to dealing with intrinsic cognitive load (i.e., germane resources).

More knowledgeable learners, on the other hand, benefit optimally from autonomous problem solving, because they have already acquired knowledge of how to solve that type of problem, which can guide their problem solving. Instructional formats that are beneficial for novice learners lose their effectiveness and can even have negative consequences for more knowledgeable learners (i.e., expertise reversal effect; Kalyuga, Ayres, Chandler, & Sweller, 2003; Kalyuga, Chandler, Tuovinen, & Sweller, 2001; Leppink, Broers, Imbos, Van der Vleuten, & Berger, 2012a, 2012b, 2013b). The information presented in worked examples is redundant for more knowledgeable learners, who are able to solve the problem without instructional guidance, and processing redundant information contributes to extraneous cognitive load (i.e., redundancy effect; Chandler & Sweller, 1991).

1.2. Measurement of cognitive load with subjective rating scales

Subjective rating scales like Paas' (1992) nine-point unidimensional mental effort rating scale have been used intensively for measuring the overall cognitive load experienced by learners (for reviews: Paas, Tuovinen, Tabbers, & Van Gerven, 2003; Van Gog & Paas, 2008). Mental effort has been defined by Paas et al., as "the cognitive capacity that is actually allocated to accommodate the demands imposed by the task; thus, it can be considered to reflect the actual cognitive load" (Paas, Tuovinen et al., 2003, p. 64; see also Paas & Van Merriënboer, 1994b). It is not entirely clear to what extent workload and cognitive load refer to the same concept across contexts, but the multidimensional NASA-TLX (Hart & Staveland, 1998) is an example of another instrument that subjectively assesses experienced workload on five seven-point rating scales. Increments of high, medium, and low estimates for each point result in 21 gradations on the scales (Hilbert & Renkl, 2009; Zumbach & Mohraz, 2008).

While measuring overall experienced cognitive load by subjective or objective techniques can be informative – especially in relation to measures of learning outcomes (Van Gog & Paas, 2008) – it is less specific than measurement of different types of cognitive load separately when it comes to informing the design of instruction. Therefore, several studies have attempted to develop instruments for measuring the three types of cognitive load separately (Ayres, 2006; Cierniak, Scheiter, & Gerjets, 2009; De Leeuw & Mayer, 2008; Eysink et al., 2009; Galy, Cariou, & Mélan, 2012). A drawback of those studies is that one or more types of cognitive load were represented by one single item. The use of multiple indicators for each of the separate types of cognitive load might yield a more precise measurement and might enable researchers to separate the types of cognitive load more clearly than the use of a single indicator for each scale. Further, when referring to one very specific instructional feature or cognitive process to measure extraneous cognitive load or germane cognitive load, a conceptual problem may arise, because the expertise reversal effect illustrates that a particular instructional feature may be associated with germane cognitive load for one learner and with extraneous cognitive load for another learner (Kalyuga et al., 2001, 2003).

1.3. A new measurement instrument for distinguishing the three types of cognitive load

Recently, a psychometric instrument was developed that took an alternative approach to the formulation of the questions for measuring different types of cognitive load (Leppink, Paas, Van der Vleuten, Van Gog, & Van Merriënboer, 2013), which may solve the problem of not being able to distinguish between different types of cognitive load at least to a certain extent. If germane cognitive load pertains to the working memory resources allocated to dealing with intrinsic cognitive load, as suggested recently by Sweller (2010) and Kalyuga (2011), it may be difficult to distinguish between germane cognitive load and intrinsic cognitive load. Although this new psychometric instrument (Leppink, Paas et al. 2013) revealed a robust three-factor structure, for a number of reasons it is not yet clear whether these three factors indeed represent the three types of cognitive load.

Firstly, the correlation between germane cognitive load and subsequent task performance in the randomized experiment was lower than expected and not statistically significant. Secondly, the set of studies presented by Leppink, Paas et al. (2013) all focused on one single context, namely that of statistics education. If the three factors indeed represent the three types of cognitive load – or stable constructs related to these types of cognitive load – one would expect these factors to come to the surface in other contexts as well. Thirdly, the experimental manipulation applied by Leppink, Paas et al. (2013) did not really lead to expected differences in any of the three factors.

1.4. The current studies

We conducted two studies using the aforementioned psychometric instrument (Leppink, Paas et al., 2013) to investigate (1) whether this instrument can help us to distinguish between intrinsic, extraneous, and germane cognitive load, (2) whether the factors obtained from that instrument can be used as predictors of task performance, and (3) how these factors are affected by the design of instruction (Study II: focusing on problem–problem, problem–example, example–problem, and example–example pairs).

2. Study I: exploratory analysis in language and statistics

In Study I, we adapted the instrument that was originally developed and tested in statistics lectures for language lessons. Table 1 presents the two versions of the questionnaire used in Study I.

Note that the language lesson version was created by drawing a parallel between vocabulary and statistical concepts, and between grammar and statistical formulas. To be able to communicate, sufficient knowledge of important concepts and definitions is vital in the statistics knowledge domain just like sufficient knowledge of vocabulary is indispensable for being able to speak a language. Besides, both grammar and formulas require knowledge and application rules. We therefore expected the two versions of the instrument – for learning statistics and for language learning – to reveal a similar three-factor pattern and comparable internal consistency values for each of the three factors. Furthermore, the statistics lecture was part of a course that was completed by an exam. We expected exam performance to be positively correlated with the factor that was supposed to represent germane cognitive load, and negatively correlated with the factors that were supposed to represent intrinsic and extraneous cognitive load. Thus, three hypotheses were tested in Study I: the two versions of the new instrument for cognitive load measurement in the language and statistics domain yield a similar three-factor pattern and comparable internal consistency values for each of the three factors (**H1**), exam performance in the statistics domain is negatively correlated with the factors that are supposed to represent intrinsic and extraneous cognitive load (**H2**), and exam performance is positively correlated with the factor that is supposed to represent germane cognitive load (**H3**).

2.1. Methods

2.1.1. Participants and materials

The language class version of the instrument presented in Table 1 was administered in a total of fourteen language classes ($n = 108$) attended by students who had chosen a language course (see Table 2 for information on language and Common European Framework of Reference [CEFR] levels; Council of Europe, 2011) either as an elective course in their study curriculum or as a separate course.

The statistics version of the instrument was administered in a lecture in an inferential statistics course for Psychology students in the first-year of the bachelor degree program ($n = 174$). The topic

Table 1
The two versions of the 'cognitive load' questionnaire used in Study I. Items 1–3 were supposed to capture intrinsic cognitive load, items 4–6 were supposed to capture extraneous cognitive load, and items 7–10 were supposed to capture germane cognitive load.

Statistics

All of the following 10 questions refer to the lecture that just finished. Please take your time to read each of the questions carefully and respond to each of the questions on the presented scale from 0 to 10, in which '0' indicates not at all the case and '10' indicates completely the case):

0 1 2 3 4 5 6 7 8 9 10

[1] The topics covered in the lecture were very complex.
[2] The lecture covered formulas that I perceived as very complex.
[3] The lecture covered concepts and definitions that I perceived as very complex.
[4] The instructions and explanations during the lecture were very unclear.
[5] The instructions and explanations during the lecture were full of unclear language.
[6] The instructions and explanations during the lecture were, in terms of learning, very ineffective.
[7] The lecture really enhanced my understanding of the topics covered.
[8] The lecture really enhanced my understanding of the formulas covered.
[9] The lecture really enhanced my knowledge of concepts and definitions.
[10] The lecture really enhanced my knowledge and understanding of the subject.

Language

All of the following 10 questions refer to the lesson that just finished. Please take your time to read each of the questions carefully and respond to each of the questions on the presented scale from 0 to 10, in which '0' indicates not at all the case and '10' indicates completely the case):

0 1 2 3 4 5 6 7 8 9 10

[1] The topics covered during the lesson were very complex.
[2] The lesson covered grammar structures that I perceived as very complex.
[3] The lesson covered vocabulary that I perceived as very complex.
[4] The instructions and explanations during the lesson were very unclear.
[5] The instructions and explanations during the lesson were full of unclear language.
[6] The instructions and explanations during the lesson were, in terms of learning, very ineffective.
[7] The lesson really enhanced my understanding of the topics covered.
[8] The lesson really enhanced my understanding of the grammar structures covered.
[9] The lesson really enhanced my knowledge of vocabulary.
[10] The lesson really enhanced my knowledge and understanding of the language.

Table 2
Language, aimed CEFR level, and number of students (of the in total 108) per language class in Study I.

Group	Language	Aimed CEFR level	Number of students
1	French	A2	7
2	Dutch	A2	9
3	Dutch	B2	6
4[a]	Italian	A1	7
5	Spanish	A2	8
6	Portuguese	A1	6
7	German	A1	8
8[a]	Italian	B1	7
9	Chinese	A1	9
10	Russian	A1	7
11	Spanish	A1	8
12	French	B1	10
13	Dutch	A1	6
14	French	B2	10

[a] The same teacher.

treated in the lecture, the sampling distribution of the sample mean (and related concepts like the standard error), formed the core of the statistics course that also covered the concepts of null hypothesis significance testing, test statistics *z*, *t*, and *F* for testing hypotheses about population means, and test statistic χ^2 for hypotheses about population proportions. All topics in the course and all questions in the exam at the end of the course were directly related to the content of this specific lecture. Students who do not understand the concepts of sampling distribution, standard error, and related concepts from sampling theory, cannot grasp the logic of null hypothesis significance testing and test statistics (Ben-Zvi & Garfield, 2004; Leppink, Broers, Imbos, Van der Vleuten, & Berger, 2011; Leppink, Broers, Imbos, Van der Vleuten, & Berger, 2013a), and are therefore likely to fail an exam that requires an understanding of these concepts.

2.1.2. Procedure

The language classes were given in small groups and no performance data could be collected. Questionnaire data were collected in different language classes taught at different CEFR levels (see Table 2) and by different teachers. While such an approach does not allow one to compare outcomes of an instrument across languages, teachers, and/or language levels – which was not the purpose of the study – it does allow for collecting data that will allow for determining the psychometric properties of the instrument. Students completed the questionnaire at the end of their two-hour class. The statistics lecture lasted two hours, and students were instructed to complete the questionnaire at the end of the lecture. All 174 students completed the questionnaire on paper at the very end of the lecture and handed it in right away. Most of these students completed the course exam five weeks later ($n = 151$).

Since Leppink, Paas et al. (2013) demonstrated that varying the order in which the items are asked does not significantly influence factor loadings or internal consistency values, we decided to present the ten items in the order presented in Table 1; note that the item order was the same for both questionnaires.

2.1.3. Data analysis

Students in the language courses were nested within learning groups. Such a multilevel design may induce a within-groups between-students correlational structure that can result in varying relationships between variables (here: the factors in the questionnaire) across learning groups. However, probably due to the limited number of both classes and students within classes, such intra-class coefficients were small to negligible and not statistically significant. Furthermore, the total sample size of the language classes ($n = 108$) was too small for confirmatory factor analysis, while following the rule of thumb that a number of participants as large as ten times the number of items (i.e., ten) in the instrument, the group was large enough for exploratory factor analysis. We therefore proceeded as follows.

Principal factor analysis was performed and internal consistency values per factor were computed for both questionnaire versions to test whether the two versions of the new instrument for cognitive load measurement in the language and statistics domain yield a similar three-factor pattern and comparable internal consistency values for each of the three factors (**H1**). Further, to test whether exam performance in the statistics domain is negatively correlated with the factors that are supposed to represent intrinsic and extraneous cognitive load (**H2**) and positively correlated with the factor that is supposed to represent germane cognitive load (**H3**), correlations were computed between the three factors derived from the statistics lecture data and course exam. These correlations are based on the data of the 151 students (86.8 percent of the 174 respondents in the lecture) who completed the course exam.

2.2. Results

Descriptive statistics revealed no abnormal response patterns or extreme cases. In both learning contexts, responses to the items covered the full range or nearly the full range from 0 to 10, and most of the items displayed absolute skewness and kurtosis values within the [−1.5; 1.5] or even [−1; 1] range. The same holds for exam score in the statistics course, which was a sum score of the number of correct responses on a total of 17 multiple-choice questions that had one correct alternative each (yielding a theoretical range from 0 to 17). The average exam score was 10.30 with a standard deviation of 2.98.

The data from both questionnaires were suitable for principal factor analysis. The values had sufficient intercorrelation, Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy was .814 for statistics and .748 for language, which are good values, and Bartlett's test of sphericity was statistically significant, $\chi^2(45) = 1223.962$, $p < .001$ for statistics, and $\chi^2(45) = 616.651$, $p < .001$ for language. Given that a three-factor solution was expected and correlations between the factors were expected (based on Leppink, Paas et al., 2013), oblique (i.e., Oblimin) rotation was performed to account for the intercorrelation of factors.

A three-factor solution was found for both questionnaires, which explained 80.8 percent of the total variance for statistics and 75.6 percent of the total variance for language. In both contexts, the items supposed to represent germane cognitive load loaded on the first factor, the items supposed to represent intrinsic cognitive load loaded on the second factor, and the items supposed to represent extraneous cognitive load loaded on the third factor. Table 3 presents the respective factor loadings for the two questionnaires and the internal consistency (i.e., Cronbach's alpha values per factor), and Table 4 presents the correlations between the three factors as a justification for oblique rotation.

The two versions yield similar factor loadings and internal consistency values. The factor correlations are similar to those reported by Leppink, Paas et al. (2013): negative between the factors supposed to represent extraneous and germane cognitive load, positive between the factors supposed to represent intrinsic and extraneous cognitive load, and around zero between the factors supposed to intrinsic and germane cognitive load. Average scores were computed per factor for each student. Table 5 presents mean, standard deviation, skewness, and kurtosis for each of the three factors per questionnaire.

In line with the hypothesis that exam performance in the statistics domain is negatively correlated with the factors that are supposed to represent intrinsic and extraneous cognitive load (**H2**), exam performance was negatively correlated with the factors

Table 3
Factor loadings and internal consistency (i.e., Cronbach's alpha) values per factor in Study I.

Factor/item	Statistics		Language	
	Loading	Alpha	Loading	Alpha
'Intrinsic'		.893		.816
Item 1	.932		.893	
Item 2	.782		.734	
Item 3	.846		.688	
'Extraneous'		.785		.838
Item 4	.693		.925	
Item 5	.911		.881	
Item 6	.569		.557	
'Germane'		.947		.889
Item 7	.922		.904	
Item 8	.858		.635	
Item 9	.905		.809	
Item 10	.933		.909	

Table 4
Correlations between the three factors in Study I.

Factor pair	Statistics	Language
'Intrinsic'–'extraneous'	.504	.271
'Intrinsic'–'germane'	−.125	.029
'Extraneous'–'germane'	−.210	−.519

supposed to represent intrinsic ($r = -.210$, $p = .010$) and extraneous ($r = -.320$, $p < .001$) cognitive load during the lecture. In line with the hypothesis that exam performance is positively correlated with the factor that is supposed to represent germane cognitive load (**H3**), exam performance was positively correlated with the factor supposed to represent germane cognitive load during the lecture ($r = .140$, $p = .084$), although this correlation was not statistically significant at the conventional $\alpha = .05$ level.

2.3. Discussion

The factor structure replication from Leppink, Paas et al. (2013) in a similar context (statistics education) and in a different context (language learning) suggests that the three factors are robust. Whether these factors represent the three types of cognitive load, however, is not yet clear. While the correlation between exam performance and the factor supposed to represent germane cognitive load was positive, it was small and not statistically significant.

Furthermore, it can be questioned what the positive correlation between the factors supposed to represent intrinsic and extraneous cognitive load means. If intrinsic, extraneous, and germane cognitive load were indeed independent and additive types of cognitive load as Sweller et al. (1998) suggest, then one would expect correlations between these types of cognitive load to be (around) zero. Seen in that light, the positive correlation between the factors supposed to represent intrinsic and extraneous load may suggest that respondents could have difficulties distinguishing between them. It may be partly for this reason why attempts to distinguish between intrinsic and extraneous (and germane) cognitive load have not been very successful until now.

On the other hand, since we did not control prior knowledge and psychology students may differ in their prior knowledge of mathematics depending on their secondary education trajectory, it might just be the case that students for whom the intrinsic cognitive load of the task was higher also experienced higher extraneous cognitive load. Moreover, the negative correlation between exam performance and the two factors supposed to represent intrinsic and extraneous cognitive load is in line with theoretical expectations and empirical findings. High extraneous (i.e., ineffective) cognitive load can be expected to hamper learning (Sweller & Chandler, 1994). Besides, it has been demonstrated that the higher the intrinsic cognitive load of a task (i.e., task complexity) the lower the learning outcomes tend to be (Ayres, 2006), and that intrinsic cognitive load is affected by individual differences in prior knowledge (Sweller, 2010; Sweller et al., 2011, 1998). As such, the negative correlation between intrinsic cognitive load and exam performance could reflect that students who generally find it difficult to learn statistical concepts (and/or mathematical formulas underlying these concepts) experienced higher intrinsic cognitive load and performed more poorly on the exam.

Table 5
Mean (M) and standard deviation (SD) along with skewness and kurtosis for each of the three factors per questionnaire in Study I.

Factor	Statistics M (SD)	Skewness	Kurtosis	Language M (SD)	Skewness	Kurtosis
'Intrinsic'	4.63 (2.03)	−.138	−.862	4.45 (1.90)	−.080	−.092
'Extraneous'	2.09 (1.55)	1.571	4.076	1.74 (1.58)	.930	−.116
'Germane'	6.30 (1.73)	−.710	1.006	7.13 (1.64)	−.699	.847

If the two factors that are supposed to represent intrinsic and extraneous cognitive load indeed represent these two types of cognitive load, this would imply that we have developed a psychometric instrument that can help researchers to differentiate intrinsic and extraneous cognitive load. However, there is another possibility we did not really anticipate in Study I (and neither did Leppink, Paas et al. 2013): responses to the items may reflect an estimation or perception of (expected) task complexity, a reflection on (expected) sources of ineffectiveness in the instruction, and a reflection on understanding and knowledge acquisition independent of any actually invested effort in these three factors and in that case independent of any aspect of working memory. The items supposed to capture intrinsic and extraneous cognitive load could reflect an estimation of the required intrinsic and extraneous cognitive load activity rather than an indication of the actually invested effort in intrinsic and extraneous cognitive load, and a similar reasoning might hold for the items supposed to capture germane cognitive load.

To investigate these possible explanations further and to examine a more direct link between the different types of cognitive load experienced and learning outcomes, an experiment was conducted in Study II in which task formats and order were varied, using a topic that participants would be novices on (Bayes' theorem), and adding the three items presented in Table 6 to the original ten items presented in Table 1.

If the three factors indeed represent intrinsic, extraneous, and germane cognitive load, one would expect the three items – which asked about actually invested mental effort for each of the factors – to contribute to the internal consistency of the factor in question.

3. Study II: problems, examples, task performance, and cognitive load

In Study II, a randomized experiment was conducted. First-year bachelor students in the Social and Health Sciences studied an application of Bayes' theorem, a difficult topic on which they were novices, in problem–problem, problem–example, example–problem or example–example condition, and a modified version of the psychometric instrument was used for cognitive load measurement after studying and after subsequent task performance.

This study also allows for replicating the findings of the study by Van Gog, Kester, and Paas (2011) in a different domain. Their study was the first to compare these four instructional conditions within one experiment, and they found some intriguing results. First of all, their study was the first to compare an example–example and example–problem condition and to demonstrate that there was no significant difference in test performance in the example–example and example–problem conditions, and no significant difference in mental effort invested in the studying phase. In other words, solving a problem after having studied worked example did not enhance test performance compared to studying another example. Van Gog and Kester (2012) replicated this finding regarding the immediate test and demonstrated that at a delayed test one week later, performance in an example–example condition was even better than in an example–problem condition. Secondly, in the study by Van Gog et al. (2011), both example–example and example–problem resulted in better test performance than problem–example and problem–problem, which, moreover, was reached with lower invested mental effort in the studying phase in the example–example and example–problem conditions than in

Table 6
The 'cognitive load' questionnaire including the three new items, where items 4 (like items 1–3 supposed to capture intrinsic cognitive load), 8 (like items 5–7 supposed to capture extraneous cognitive load), and 13 (like items 9–12 supposed to capture germane cognitive load) were new in Study II.

All of the following 10 questions refer to the activity that just finished. Please take your time to read each of the questions carefully and respond to each of the questions on the presented scale from 0 to 10, in which '0' indicates not at all the case and '10' indicates completely the case):
0 1 2 3 4 5 6 7 8 9 10
[1] The content of this activity was very complex.
[2] The problem/s covered in this activity was/were very complex.
[3] In this activity, very complex terms were mentioned.
[4] I invested a very high mental effort in the complexity of this activity.
[5] The explanations and instructions in this activity were very unclear.
[6] The explanations and instructions in this activity were full of unclear language.
[7] The explanations and instructions in this activity were, in terms of learning, very ineffective.
[8] I invested a very high mental effort in unclear and ineffective explanations and instructions in this activity.
[9] This activity really enhanced my understanding of the content that was covered.
[10] This activity really enhanced my understanding of the problem/s that was/were covered.
[11] This activity really enhanced my knowledge of the terms that were mentioned.
[12] This activity really enhanced my knowledge and understanding of how to deal with the problem/s covered.
[13] I invested a very high mental effort during this activity in enhancing my knowledge and understanding.

the problem–example and problem–problem conditions. This finding is in line with a large number of studies on the worked example effect (for reviews, see Sweller et al., 1998; Van Gog & Rummel, 2010). Thirdly, the problem–example and problem–problem conditions did not differ from each other in test performance or invested mental effort in the studying phase.

At first, the finding that the example–problem condition performed better than the problem–example condition may be surprising. After all, students in these two conditions received the same amount of instructional support, only in a different condition. However, this finding is also in line with findings by Reisslein, Atkinson, Seeling, and Reisslein (2006), and suggests that condition matters because studying an example first allows for building a cognitive schema (i.e., germane cognitive load activity) that can subsequently be used when solving the problem (lowering the intrinsic cognitive load of the problem compared to a problem–problem condition). When solving a problem first, there may be high extraneous cognitive load and little learning from solving that problem. However, the latter does not yet explain why the problem–example condition did not improve performance beyond that in the problem–problem condition. Again, the Van Gog et al. (2011) study was the first to compare these two conditions, and this finding was unexpected. A possible explanation provided by Van Gog et al. is that students get so frustrated by the problem-solving experience that they do not study the subsequent example very well: something which should be reflected in differences in germane cognitive load ratings between the example–problem condition and the problem–example condition if we could measure those. Therefore, it would not only be interesting to see if these findings could be replicated in another domain, but also to measure experienced cognitive load with the new instrument, to determine whether we can better explain this pattern of findings.

In line with these findings, we expected our novice participants who studied the application of Bayes' theorem in the example–problem or example–example condition to perform better on a subsequent posttest on applications of Bayes' theorem than participants who studied this application in the problem–example or problem–problem condition (**H4a**), but that the second format did not influence posttest performance significantly (**H4b**). In terms of cognitive load, we hypothesized that the three items added to the cognitive load instrument – asking about actually invested mental effort for each of the factors – would contribute to the internal consistency of the factors supposed to represent intrinsic cognitive load (**H5a**), extraneous cognitive load (**H5b**), and germane cognitive load (**H5c**). Further, we expected to find that students in the example–example and example–problem condition would report lower intrinsic cognitive load on the posttest (**H6**), lower extraneous cognitive load in the studying phase as well as on the posttest (**H7a** and **H7b**), and higher germane cognitive load in the studying phase and on the posttest (**H8a** and **H8b**) than students in the problem–example and problem–problem condition.

3.1. Methods

3.1.1. Participants and experimental design

A total of eighty-four first-year bachelor students in the Social and Health Sciences were allocated randomly to the problem–problem, problem–example, example–problem, and example–example condition. The first task entailed either autonomous problem solving or the study of a worked example, and the same holds for the second task. To account for potential context effects, we designed two problems in a different context. The two contexts were presented in counterbalanced order, meaning that about half of the participants completed their first task in one context while the others completed their first task in the other context, and vice versa. Nine students canceled their participation last minute due to changes in their educational timetable, and another one other student did not comply with the instructions and was therefore excluded from the experiment. This resulted in the following situation: problem–problem ($n = 20$), problem–example ($n = 18$), example–problem ($n = 18$), and example–example ($n = 18$).

3.1.2. Materials and procedure

3.1.2.1. Cognitive load instrument. We slightly reformulated some of the items and added three items to the instrument for cognitive load measurement in Study II to obtain a better understanding of what the three factors actually represent (see Table 6).

3.1.2.2. Learning and test materials. The two problems (i.e., contexts) focused on the same application of Bayes' theorem:

$$P(A|B) = [P(A) \times P(B|A)]/P(B)$$

There were two reasons why this topic was chosen. Firstly, all participants would study this topic in a subsequent statistics course in their curriculum. A one-hour lecture on applications of this theorem followed immediately after participation in the experiment, providing a realistic educational context in which participants were motivated to learn. Secondly, statistics is an important tool for scientific research as well as in other professions and even in daily life. Statistics is largely about mathematical modeling of empirical phenomena, it is the science of uncertainty, and conditional probabilities form the core of all statistics.

3.1.3. Procedure

The procedure was a follows. The experiment lasted half an hour. Participants were given five minutes to do the first task (either autonomous problem–solving or example in either first or second context) on paper, and – after handing in the first task – five minutes for the second task (either autonomous problem–solving or example in either first or second context). Both tasks required a

calculation that would result in a conditional probability and as such could be viewed as open-ended questions (no multiple-choice alternatives were provided). Participants who had to perform the task autonomously had to provide this conditional probability themselves, while participants who studied the worked example saw the correct calculation and conditional probability as solution to the problem.

After these first ten minutes, students completed the adjusted 'cognitive load' questionnaire (thirteen items) for the first time, rating the cognitive load experienced in this ten-minutes training phase. After a five-minutes break, students received a posttest consisting of six open-ended questions built around two new contexts (i.e., three questions around each context) but requiring exactly the same application of Bayes' theorem. The questions followed exactly the same story line as the problems that required autonomous problem solving in the studying phase, meaning that participants had to calculate and provide the correct conditional probability by themselves. The contexts and questions were made such that it was impossible to provide the correct conditional probability through the use of an incorrect algorithm (e.g., computing the joint probability $P(A, B)$ or providing conditional probability $P(B|A)$ instead of the correct $P(A|B)$) and that it was virtually impossible to just guess the correct answer. This way, one can be rather confident that a correct response reflects that the participant is able to apply Bayes' theorem correctly. It is therefore of little surprise that the sum of correct responses (i.e., an integer on the scale of 0–6) yielded a very high Cronbach's alpha ($\alpha = .950$) with all item p-values (i.e., percentage of correct response) in the range of .70–.78 and corrected item–total correlations of .695 and higher. This yielded a very precise measurement of the extent to which a student is able to apply Bayes' theorem, and thus optimal statistical power for treatment effects of instructional conditions on participants' posttest performance.

All students managed to complete all six questions within the fifteen minutes they were given for the posttest. Having completed the posttest, all students completed the adjusted 'cognitive load' questionnaire (thirteen items) once again, rating the cognitive load experienced in the posttest.

3.1.4. Data analysis

To examine the effects of instructional conditions (i.e., the four conditions) on posttest performance, we performed two-way between-subjects (BS) analysis of variance (ANOVA), because this enables one to test three specific contrasts: (1) first task autonomous problem solving vs. worked example (i.e., main effect of type of first task), (2) second task autonomous problem solving vs. worked example (i.e., main effect of type of second task), and (3) an extra effect of one specific condition (i.e., interaction effect of types first and second task). Expanding BS ANOVA to three-way to account for counterbalanced training context to our model did not contribute to the explanation of posttest performance; counterbalanced training context had only a small and not statistically significant effect on posttest performance, $F(1, 72) = .891$, $p = .348$, $\eta^2 = .012$ (values of .01, .06, and .14 are indicative of a small, medium, and large effect, respectively; Field, 2013). Thus, the two-way BS ANOVA sufficed. We expected a statistically significant main effect of first format (**H4a**), no statistically significant main effect of second format (**H4b**), and no statistically significant interaction effect.

To examine the added value of the additional three items in the 'cognitive load' questionnaire (**H5a**, **H5b**, and **H5c**), we computed per factor and per measurement occasion (i.e., training and posttest) Cronbach's alpha for each of three factors along with corrected item–total correlations and the Cronbach's alpha value in the case of deletion of a particular item. We then computed average scores per factor per measurement occasion to investigate the correlations between posttest performance and the (averaged) factor scores per measurement occasion. To test the hypotheses that students starting with a worked example would report lower intrinsic cognitive load on the posttest (**H6**), lower extraneous cognitive load in the studying phase (**H7a**) as well as on the posttest (**H7b**), and higher germane cognitive load in the studying phase (**H8a**) and on the posttest (**H8b**), we performed split-plot ANOVA, using the same BS factors (first and second task) as in the two-way ANOVA and treating the scores for each factor from the two measurement occasions (training and posttest) as repeated measures (i.e., within-subjects, WS). The latter enables testing for BS by WS interaction (which is present if some conditions differ in intrinsic cognitive load on the posttest but not in the studying phase) and yields more power for testing main effects of first and second format than testing these main effects for studying phase and posttest performance separately.

Finally, as was the case for posttest performance, including counterbalanced training context in the analysis did not contribute to the explanation of any of the factor scores; all η^2-values were in the range of .005 and .014 (comparable to the η^2-value of .012 in the case of posttest performance) and p-values ranged from .309 to .542.

3.2. Results

Posttest performance was somewhat skewed to the left (skewness = −1.195), with a mean score (M) of 4.49 and a standard deviation (SD) 2.34. The problem–problem condition performed worst ($M = 3.50$, $SD = 2.78$), followed by the problem–example condition ($M = 4.11$, $SD = 2.52$), followed by the example–problem condition ($M = 5.06$, $SD = 1.83$), and the example–example condition performed best ($M = 5.39$, $SD = 1.65$). As expected, the interaction effect between first and second task was not statistically significant, $F(1, 70) = .070$, $p = .793$, $\eta^2 = .001$. This means that effects of first and second task – if existing – can be viewed as additive.

Although participants who studied a worked example in the second task (i.e., problem–example or example–example) scored on average .472 points (range: 0–6) higher than participants who solved a problem autonomously in the second task, this difference was not statistically significant, $F(1, 70) = .805$, $p = .373$, $\eta^2 = .011$, and the effect size indicates that we are talking about a small effect at best. The effect of the first task, however, was statistically significant, $F(1, 70) = 7.245$, $p = .009$, $\eta^2 = .094$, and the effect size indicates a medium to somewhat larger effect; participants who studied a worked example in the first task performed much better on the posttest (1.417 points on the 0–6 scale) than participants who solved a problem autonomously in the first task.

Table 7 presents mean (M) and standard deviation (SD) along with skewness and kurtosis of the (averaged) scores of the factors supposed to represent intrinsic, extraneous, and germane cognitive load per measurement occasion (training and posttest) per condition.

Table 8 presents Cronbach's alpha values along with corrected item–total correlations and Cronbach's alpha values after deletion of a particular item (keeping the rest of the items) per factor per measurement occasion, and Table 9 presents correlations with posttest performance.

Table 8 indicates that the three factors yielded comparable Cronbach's alpha values for training and posttest, only a somewhat lower Cronbach's alpha value for the factor supposed to represent extraneous cognitive load for the training which may reflect a restriction of range effect. Table 9 indicates that the correlations between posttest performance and the three factors supposed to

Table 7
Mean (*M*) and standard deviation (*SD*) along with skewness and kurtosis of the (averaged) scores of the factors supposed to represent intrinsic, extraneous, and germane cognitive load per measurement occasion (training and posttest) per instructional format (Study II).

Factor	Training *M* (*SD*)	Skewness	Kurtosis	Posttest *M* (*SD*)	Skewness	Kurtosis
Problem–problem						
'Intrinsic'	2.03 (1.44)	.646	−.575	2.26 (1.83)	.786	.151
'Extraneous'	1.09 (1.36)	1.284	.550	1.49 (1.76)	1.095	−.234
'Germane'	2.03 (1.84)	.487	−1.267	1.94 (1.74)	.599	−.291
Problem–example						
'Intrinsic'	1.78 (1.61)	.749	−.537	3.10 (1.97)	.608	1.105
'Extraneous'	1.14 (1.30)	1.180	1.025	1.43 (1.48)	1.264	−.413
'Germane'	1.74 (1.47)	.361	−.994	2.42 (1.98)	.678	−.942
Example–problem						
'Intrinsic'	2.60 (1.85)	.584	−.623	2.65 (1.47)	.498	−.913
'Extraneous'	1.54 (1.49)	.772	−.782	1.86 (1.47)	.503	−.048
'Germane'	4.12 (2.43)	−.125	−1.305	4.56 (2.11)	−.247	−.533
Example–example						
'Intrinsic'	2.32 (1.62)	.046	−.778	2.79 (1.89)	.392	−.767
'Extraneous'	2.01 (1.72)	.870	.007	2.07 (1.74)	.642	−.959
'Germane'	2.72 (2.09)	.526	−.566	3.00 (1.73)	.178	.200

represent intrinsic, extraneous, and germane cognitive load are similar for the two measurement occasions and (especially for the first two factors) close to zero.

Finally, Tables 10–12 present the outcomes of split-plot ANOVA for the three factors that were supposed to represent intrinsic, extraneous, and germane cognitive load, respectively.

None of Tables 10–12 indicate statistically significant interaction effects between format and time or between first and second format. Only first format in Table 12 is statistically significant, and the η^2-value of .170 indicates a large effect.

3.3. Discussion

In line with the study by Van Gog et al. (2011), we found that example–example and example–problem pairs led to better posttest performance than problem–problem and problem–example pairs (**H4a**), and that the second format did not influence posttest performance significantly (**H4b**).

The findings presented in Table 8 provide support for the hypothesis that two of the three items added to the cognitive load instrument – asking about actually invested mental effort for each of the factors – would contribute to the internal consistency of the factors supposed to represent intrinsic cognitive load (**H5a**) and extraneous cognitive load (**H5b**). With regard to these two factors, the corrected item–total correlations (and Cronbach's alpha values without the item in question) indicate that the added items – asking about actually invested mental effort (in intrinsic or extraneous features) – contribute to the internal consistency. This appears to indicate that the items supposed to capture intrinsic cognitive load measure one and the same latent construct, and the items supposed to capture extraneous cognitive load measure one and the same latent construct. The two constructs in question may be intrinsic and extraneous cognitive load or factors related to these two types of cognitive load.

For the third factor, which was supposed to represent germane cognitive load, the findings appear less convincing. The findings presented in Table 12 provide support for the hypothesis that students in the example–example and example–problem condition would report higher germane cognitive load in the studying phase (**H8a**) and on the posttest (**H8b**) than students in the problem–example and problem–problem condition (the η^2-value of .170 indicating a large effect). However, the findings presented in Table 8 do not provide support for the hypothesis that the item asking about actually invested mental effort in germane cognitive load activity would contribute to the internal consistency of the factor supposed to represent germane cognitive load (**H5c**). While the corrected item–total correlation of the added item is still quite high, this value is distinguishable from the values of the other items, which may reflect that the other four items do not directly capture actually invested effort in germane cognitive load activity. This may explain the modest correlations between the factor supposed to represent germane cognitive load activity and posttest performance in Study I and in the experiment by Leppink, Paas et al. (2013), which are not any higher in Study II (see Table 9). While these correlations are again positive, their magnitude appears to indicate that the relation between the third factor and germane cognitive load is limited at best.

Table 8
Cronbach's alpha values along with corrected item-total correlations and Cronbach's alpha values after deletion of a particular item (keeping the rest of the items) per factor per measurement occasion, and correlations with posttest performance (Study II).

Factor/item	Training			Posttest		
	Corrected item–total correlation	Cronbach's alpha		Corrected item–total correlation	Cronbach's alpha	
		Factor	Item out		Factor	Item out
'Intrinsic'		.853			.872	
Item 1	.849		.740	.833		.790
Item 2	.819		.755	.876		.770
Item 3	.493		.886	.465		.921
Item 4	.662		.826	.764		.821
'Extraneous'		.632			.787	
Item 5	.451		.537	.725		.661
Item 6	.525		.562	.622		.755
Item 7	.363		.631	.431		.837
Item 8	.469		.531	.734		.663
'Germane'		.933			.931	
Item 9	.921		.897	.894		.899
Item 10	.892		.903	.891		.900
Item 11	.826		.917	.834		.913
Item 12	.802		.922	.848		.909
Item 13	.680		.942	.637		.948

Table 9
Correlations between the three factors per measurement occasion and posttest performance (Study II).

Factor	Training *r* (*p*-value)	Posttest *r* (*p*-value)
'Intrinsic'	.068 (.567)	−.017 (.883)
'Extraneous'	.049 (.677)	.042 (.722)
'Germane'	.141 (.230)	.113 (.336)

Table 10
Outcomes of split-plot ANOVA for the factor that was supposed to represent intrinsic cognitive load (Study II).

Effect	$F(1, 70)$	p-value	η^2-value
Time	7.121	.009	.092
First task	.731	.395	.010
Second task	.101	.751	.001
First task *by* second task	1.204	.607	.004
Time *by* first task	1.728	.193	.024
Time *by* second task	3.715	.058	.050
Time *by* first task *by* second task	.737	.393	.010

Table 12
Outcomes of split-plot ANOVA for the factor that was supposed to represent germane cognitive load (Study II).

Effect	$F(1, 70)$	p-value	η^2-value
Time	3.251	.076	.044
First task	14.337	<.001	.170
Second task	2.782	.100	.038
First task *by* second task	3.632	.061	.049
Time *by* first task	.029	.865	<.001
Time *by* second task	.722	.398	.010
Time *by* first task *by* second task	1.643	.204	.023

The finding that the different conditions did not differ in average score on the factors supposed to represent intrinsic and extraneous cognitive load was unexpected. It is not in line with the hypothesis that students in the example–example and example–problem condition would report lower intrinsic cognitive load on the posttest (**H6**) or with the hypothesis that students in the example–example and example–problem condition would report lower extraneous cognitive load in the studying phase (**H7a**) as well as on the posttest (**H7b**) than students in the problem–example and problem–problem conditions. Possibly, the learning phase was too short to significantly affect intrinsic or extraneous cognitive load, or perhaps the beneficial effects of worked examples are rather due to germane cognitive load. However, in this light it is perhaps even more intriguing that the findings presented in Table 12 do provide support for the hypothesis that students in the example–example and example–problem condition would report higher germane cognitive load in the studying phase and on the posttest than students in the problem–example and problem–problem condition. It may indicate that students are able to rate this 'knowledge and understanding' factor and judge it differently between conditions. If so, even if not germane load as such, it is still a potentially important construct.

4. General discussion

Together, the findings appear to provide some support for the assumption that intrinsic and extraneous cognitive load can be differentiated using a psychometric instrument, or at least that the factors that were supposed to represent intrinsic and extraneous cognitive load relate to intrinsic and extraneous cognitive load. The two factors appear consistently across studies, newly added items asking about actually invested effort contribute to the reliability of each of the two factors, and the questions loading on these factors can be related to the theoretical concepts of intrinsic and extraneous cognitive load. Although these findings appear to provide support for the assumption that the two factors either represent or closely relate to intrinsic and extraneous cognitive load, respectively, further experimentation is needed to examine the validity of this assumption.

We recommend using the two factors supposed to represent intrinsic and extraneous cognitive load, as used in Study II, in a series of new experiments that, as in Study II, replicate known effects. It would be interesting to examine how these factors behave in conditions of which we can be very confident that they impose very high intrinsic and/or extraneous cognitive load and to see how these factors correlate with task performance in such conditions. Especially the fact that no significant difference in extraneous cognitive load during training was found is surprising and needs further study. On the one hand, this might suggest that our instrument would not actually measure extraneous cognitive load. On the other hand, it is possible that the theoretical explanation of the worked example effect lies more in germane cognitive load than in extraneous cognitive load effects. Also, to investigate the possibility that the learning phase was too short to significantly affect intrinsic or extraneous cognitive load, replications of Study II should use a longer learning phase.

Table 11
Outcomes of split-plot ANOVA for the factor that was supposed to represent extraneous cognitive load (Study II).

Effect	$F(1, 70)$	p-value	η^2-value
Time	2.635	.109	.036
First task	3.323	.073	.045
Second task	.276	.601	.004
First task *by* second task	.285	.595	.004
Time *by* first task	.232	.631	.003
Time *by* second task	.321	.573	.005
Time *by* first task *by* second task	.056	.814	.001

The fact that the two factors supposed to represent intrinsic and extraneous cognitive load are correlated appears to reflect that students have some difficulties distinguishing the two types of cognitive load. Part of the problem may lie in specific question wording effects. Unclear instruction may not necessarily result from additional and irrelevant processing but at least to some extent from a lack of prior knowledge, and instruction may be complex to a learner because it involves many cognitive activities some of which could be irrelevant. It is worth testing wording effects like these in future experiments.

The findings appear to be in line with the recently proposed reconceptualization of germane cognitive load as referring to the actual working memory resources devoted to dealing with intrinsic cognitive load (Kalyuga, 2011; Sweller, 2010; Sweller et al., 2011). At least, the findings do not point against such a reconceptualization. Support for the assumption that the third factor in the psychometric instrument represents or closely relates to germane cognitive load is limited. The added item on invested effort did not contribute to the reliability of this factor, which does not support the assumption that the items originally assumed to capture germane cognitive load actually refer to a particular type of allocated working memory resources.

The findings of Study II may indicate that students can rate the 'knowledge and understanding' factor and judge it differently across conditions. If so, even if not germane load as such, it is still a potentially important construct. One explanation for the consistently small correlation between this factor and performance may be that learners can expend much or little effort on what they perceive to be learning but part of that effort may be fruitless. A second explanation in this context may be that without additional feedback students can reflect on their relevance of effort only to limited extent and perhaps only more advanced students or experts in a domain really succeed in doing so. The latter is worth investigating in new studies which include novices and experts in a variety of domains.

The finding that starting self-study of a new topic with a worked example instead of with autonomous problem solving has beneficial effects on task performance is in line with previous research on worked example effects (Cooper & Sweller, 1987; Paas, 1992; Paas & Van Merriënboer, 1994a; Sweller & Cooper, 1985; Van Gog et al., 2006) and with previous research on problem–example and

example problem pairs (Reisslein et al., 2006; Van Gog et al., 2011). This finding is particularly interesting in the light of recent debates on instructional guidance (e.g., Clark, Kirschner, & Sweller, 2012), as it illustrates that it not only matters how much guidance is provided, but also when it is provided. It would be interesting to replicate Study II with a different topic and with both novices and more advanced learners, to examine (1) whether problem–problem and/or problem–example pairs become more effective relative to example–problem and example–example pairs as learners' proficiency increases (which would be in line with findings on the expertise reversal effect; Kalyuga et al., 2001, 2003; Leppink, Broers et al., 2012a, 2012b, 2013b), and (2) how the factors supposed to represent intrinsic and extraneous cognitive load behave for different levels of expertise in different instructional conditions.

Although law cases and empirical research often appear to have little in common, at least one thing they do have in common: validity of a story is about a chain of evidence (Kane, 2006). Whether we deal with pieces of evidence in a law case or with empirical studies, we make assumptions, and absolute proof does not exist. Like a suspect in a criminal case should not be convicted based on only one piece of evidence – *unus testis nullus testis* – validity of a measurement instrument is not established in one or two (sets of) studies; it is a journey in search for a chain of evidence, and to obtain that chain of evidence some elements in the instrument may need revision or adjustment. The recent suggestion to redefine germane cognitive load activity as the working memory resources allocated to dealing with intrinsic cognitive load appears very attractive in terms of transparency and parsimony of cognitive load theory. Taken the findings of Leppink, Paas et al. (2013) and the findings of the current two studies together, there appears to be empirical support for this move. Further development of the measurement of intrinsic and extraneous cognitive should be driven by an ongoing dialog between cognitive load theory as defined now (Kalyuga, 2011; Sweller, 2010; Sweller et al., 2011) and empirical research for which some suggestions are provided in this paper.

References

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic load within problems. *Learning and Instruction, 16*, 389–400.

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 570–585. http://dx.doi.org/10.1037/0278-7393.33.3.570.

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

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*, 293–332. http://dx.doi.org/10.1207/s1532690xci0804_2.

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior, 25*, 315–324. http://dx.doi.org/10.1016/j.chb.2008.12.020.

Clark, R. E., Kirschner, P. A., & Sweller, J. (2012). Putting students on the path to learning: the case for fully guided instruction. *American Educator, 36*, 6–11. http://www.aft.org/pdfs/americaneducator/spring2012/Clark.pdf.

Cooper, G. A., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology, 79*, 347–362. http://dx.doi.org/10.1037/0022-0663.79.4.347.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 152–153.

De Leeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100*, 223–234. http://dx.doi.org/10.1037/0022-0663.100.1.223.

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. http://dx.doi.org/10.3102/0002831209340235.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). London: Sage.

Galy, E., Cariou, M., & Mélan, C. (2012). What is the relationship between mental workload factors and cognitive load types? *International Journal of Psychophysiology, 83*, 269–275. http://dx.doi.org/10.1016/j.ijpsycho.2011.09.023.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In P. A. Hancock, & N. Meshtaki (Eds.), *Human mental workload* (pp. 139–183). Amsterdam, the Netherlands: North-Holland.

Hilbert, T. S., & Renkl, A. (2009). Learning how to use a computer-based concept-mapping tool: self-explaining examples helps. *Computers in Human Behavior, 25*, 267–274. http://dx.doi.org/10.1016/j.chb.2008.12.006.

Kalyuga, S. (2011). Cognitive load theory: how many types of load does it really need? *Educational Psychology Review, 23*, 1–19. http://dx.doi.org/10.1007/s10648-010-9150-7.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*, 23–31. http://dx.doi.org/10.1207/S15326985EP3801_4.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem-solving is superior to studying worked examples. *Journal of Educational Psychology, 93*, 579–588. http://dx.doi.org/10.1037/0022-0663.93.3.579.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17–64). Westport: ACE/'Praeger.

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*. online http://www.amstat.org/publications/jse/v19n1/leppink.pdf.

Leppink, J., Broers, N. J., Imbos, Tj., Van der Vleuten, C. P. M., & Berger, M. P. F. (2012a). Prior knowledge moderates instructional effects on conceptual understanding of statistics. *Educational Research and Evaluation, 18*, 37–51. http://dx.doi.org/10.1080/13803611.2011.640873.

Leppink, J., Broers, N. J., Imbos, Tj., Van der Vleuten, C. P. M., & Berger, M. P. F. (2012b). Self-explanation in the domain of statistics: an expertise reversal effect. *Higher Education, 63*, 771–785. http://dx.doi.org/10.1007/s10734-011-9476-1.

Leppink, J., Broers, N. J., Imbos, Tj., Van der Vleuten, C. P. M., & Berger, M. P. F. (2013a). The effectiveness of propositional manipulation as a lecturing method in the statistics knowledge domain. *Instructional Science*. http://dx.doi.org/10.1007/s11251-013-9268-3.

Leppink, J., Broers, N. J., Imbos, Tj., Van der Vleuten, C. P. M., & Berger, M. P. F. (2013b). The effect of guidance in problem-based learning of statistics. *Journal of Experimental Education*. http://dx.doi.org/10.1080/00220973.2013.813365.

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*. http://dx.doi.org/10.3758/s13428-013-0334-1.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *The Psychological Review, 63*, 81–97. http://dx.doi.org/10.1037/0033-295X.101.2.343.

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. http://dx.doi.org/10.1037/0022-0663.84.4.429.

Paas, F., Tuovinen, J., Tabbers, H., & Van Gerven, P. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*, 63–71. http://dx.doi.org/10.1207/S15326985EP3801_8.

Paas, F., & Van Merriënboer, J. J. G. (1994a). Variability of worked examples and transfer of geometrical problem-solving skills: a cognate-load approach. *Journal of Educational Psychology, 86*, 122–133. http://dx.doi.org/10.1037/0022-0663.86.1.122.

Paas, F., & Van Merriënboer, J. J. G. (1994b). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6*, 51–71. http://dx.doi.org/10.1007/BF02213420.

Reisslein, J., Atkinson, R. K., Seeling, P., & Reisslein, M. (2006). Encountering the expertise reversal effect with a computer-based environment on electrical circuit analysis. *Learning and Instruction, 16*, 92–103. http://dx.doi.org/10.1016/j.learninstruc.2006.02.008.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*, 295–312. http://dx.doi.org/10.1016/0959-4752(94)90003-5.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*, 123–138. http://dx.doi.org/10.1007/s10648-010-9128-5.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York: Springer.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction, 12*, 185–223. http://dx.doi.org/10.1207/s1532690xci1203_1.

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load as a factor in the structuring of technical material. *Journal of Experimental Psychology, 119*, 176–192. http://dx.doi.org/10.1037/0096-3445.119.2.176.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*, 59–89. http://dx.doi.org/10.1207/s1532690xci0201_3.

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296. http://dx.doi.org/10.1023/A:1022193728205.

Van Gog, T., & Kester, L. (2012). A test of the testing effect: acquiring problem-solving skills from worked examples. *Cognitive Science, 36*, 1532–1541. http://dx.doi.org/10.1111/cogs.12002.

Van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example–problem, and problem–example pairs on novices' learning. *Contemporary Educational Psychology, 36*, 212–218. http://dx.doi.org/10.1016/j.cedpsych.2010.10.004.

Van Gog, T., & Paas, F. (2008). Instructional efficiency: revisiting the original construct in educational research. *Educational Psychologist, 43*, 16–26. http://dx.doi.org/10.1080/00461520701756248.

Van Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2006). Effects of process-oriented worked examples on troubleshooting transfer performance. *Learning and Instruction, 16*, 154–164. http://dx.doi.org/10.1016/j.learninstruc.2006.02.003.

Van Gog, T., & Rummel, N. (2010). Example-based learning: integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review, 22*, 155–174. http://dx.doi.org/10.1007/s10648-010-9134-7.

Van Merriënboer, J. J. G. (1990). Strategies for programming instruction in high school: program completion vs. program generation. *Journal of Educational Computing Research, 6*, 265–285. http://dx.doi.org/10.2190/4NK5-17L7-TWQV-1EHL.

Van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: recent developments and future directions. *Educational Psychology Review, 17*, 147–177. http://dx.doi.org/10.1007/s10648-005-3951-0.

Van Merriënboer, J. J. G., & Sweller, J. (2010). Cognitive load theory in health professions education: design principles and strategies. *Medical Education, 44*, 85–93. http://dx.doi.org/10.1111/j.1365-2923.2009.03498.x.

Zumbach, J., & Mohraz, M. (2008). Cognitive load in hypermedia reading comprehension: influence of text type and linearity. *Computers in Human Behavior, 24*, 875–887. http://dx.doi.org/10.1016/j.chb.2007.02.015.