Behav Res (2013) 45:1058–1072
DOI 10.3758/s13428-013-0334-1

Development of an instrument for measuring different types of cognitive load

Jimmie Leppink · Fred Paas · Cees P. M. Van der Vleuten · Tamara Van Gog · Jeroen J. G. Van Merriënboer

Published online: 10 April 2013

Abstract According to cognitive load theory, instructions can impose three types of cognitive load on the learner: intrinsic load, extraneous load, and germane load. Proper measurement of the different types of cognitive load can help us understand why the effectiveness and efficiency of learning environments may differ as a function of instructional formats and learner characteristics. In this article, we present a ten-item instrument for the measurement of the three types of cognitive load. Principal component analysis on data from a lecture in statistics for PhD students (n = 56) in psychology and health sciences revealed a three-component solution, consistent with the types of load that the different items were intended to measure. This solution was confirmed by a confirmatory factor analysis of data from three lectures in statistics for different cohorts of bachelor students in the social and health sciences (ns = 171, 136, and 148), and received further support from a randomized experiment with university freshmen in the health sciences (n = 58).

Keywords Cognitive load · Intrinsic load · Extraneous load · Germane load · Subjective rating scales

J. Leppink (✉) · C. P. M. Van der Vleuten · J. J. G. Van Merriënboer
Department of Educational Development and Research, Maastricht University, P. O. Box 616, 6200 MD Maastricht, The Netherlands
e-mail: jimmie.leppink@maastrichtuniversity.nl

F. Paas · T. Van Gog
Institute of Psychology, Erasmus University Rotterdam, Rotterdam, The Netherlands

F. Paas
Interdisciplinary Educational Research Institute, University of Wollongong, Wollongong, New South Wales, Australia

According to cognitive load theory (Sweller, 2010; Sweller, Van Merriënboer, & Paas, 1998; Van Merriënboer & Sweller, 2005), instruction can impose three types of cognitive load (CL) on a learner's cognitive system: task complexity and the learner's prior knowledge determine the intrinsic load (IL), instructional features that are not beneficial for learning contribute to extraneous load (EL), and instructional features that are beneficial for learning contribute to germane load (GL). IL should be optimized in instructional design by selecting learning tasks that match learners' prior knowledge (Kalyuga, 2009), whereas EL should be minimized to reduce ineffective load (Kalyuga & Hanham, 2011; Paas, Renkl, & Sweller, 2003) and to allow learners to engage in activities imposing GL (Van Merriënboer & Sweller, 2005).

The extent to which instructional features contribute to EL or GL may depend on the individual learner and the extent to which the individual learner experiences IL. For example, less knowledgeable learners may learn better from worked examples (i.e., worked example effect; Cooper & Sweller, 1987; Paas & Van Merriënboer, 1994; Sweller & Cooper, 1985) or from completing a partially solved problem (i.e., a problem completion effect; Paas, 1992; Van Merriënboer, 1990) than from autonomous problem-solving. More knowledgeable learners benefit optimally from autonomous problem-solving (i.e., expertise reversal effect; Kalyuga, Ayres, Chandler, & Sweller, 2003; Kalyuga, Chandler, Tuovinen, & Sweller, 2001). The information presented in worked examples is redundant for more knowledgeable learners who have the cognitive schemata to solve the problem without instructional guidance, and processing redundant information leads to EL (i.e., a redundancy effect; Chandler & Sweller, 1991). Also, when instructions are presented in such a way that learners need to split their attention between two or more mutually referring information sources they are

likely to experience higher EL (i.e., split-attention effect; Sweller, Chandler, Tierney, & Cooper, 1990).

When IL is optimal and EL is low, learners can engage in knowledge elaboration processes (Kalyuga, 2009) like self-explanation (Atkinson, Renkl, & Merrill, 2003; Berthold & Renkl, 2009) and argumentation (Fischer, 2002; Knipfer, Mayr, Zahn, Schwan, & Hesse, 2009) that impose GL and facilitate learning.

Being able to properly measure the different types of CL would help educational researchers and instructional designers to better understand why learning outcomes attained with instructional formats may differ between formats or between learners. If IL differs between learners who are given the same instructions, the difference in IL provides us with information on the learners' level of expertise and—if measured repeatedly—how that changes over time. Meanwhile, when instructions are varied—for example in experimental studies—such measurements can help us gain a better understanding of instructional effects for learners with similar or distinct levels of expertise. Thus far, however, only a few attempts have been made to develop instruments for measuring these different types of cognitive load (Cierniak, Scheiter, & Gerjets, 2009; DeLeeuw & Mayer, 2008; Eysink et al., 2009).

The measurement of CL, IL, EL, and GL

Subjective rating scales like Paas's (1992) nine-point mental effort rating scale have been used intensively (for reviews, see Paas, Tuovinen, Tabbers, & Van Gerven, 2003; Van Gog & Paas, 2008) and have been identified as reliable and valid estimators of overall CL (Ayres, 2006; Paas, Ayres, & Pachman, 2008; Paas, Tuovinen, et al., 2003; Paas, Van Merriënboer, & Adam, 1994). From the reviews by Paas, Tuovinen, et al. and Van Gog and Paas, it also becomes clear that in many studies task difficulty rather than mental effort is used as an estimator of CL. Next to measures of overall CL attempts have been made to measure the different types of CL separately. Ayres, for instance, presented a rating scale for the measurement of IL, and other researchers have used rating scales for measuring IL, EL, and GL separately (e.g., Eysink et al., 2009). To measure EL, Cierniak et al. (2009) asked learners to rate on a six-point scale how difficult it was to learn with the material, and to measure GL, they adopted Salomon's (1984) question of how much learners concentrated during learning.

Generally, the fact that different scales, varying in both number of categories and labels, are used is a problem, especially because some of these scales have not been validated. Moreover, whether overall CL or (one of) the types of CL is measured, in most cases one Likert item is used, and the number of categories in the item typically varies (see also Van Gog & Paas, 2008) and can be five (e.g., Camp, Paas, Rikers, & Van Merriënboer, 2001; Salden, Paas, Broers, & Van Merriënboer, 2004), six (e.g., Cierniak et al., 2009), seven (e.g., Ayres, 2006), or nine (e.g., Eysink et al., 2009; Paas, 1992). Although load data are typically assumed to be measured at interval level (i.e., metric), by using less than seven categories one may be measuring at ordinal level of measurement rather than at interval level of measurement. Furthermore, when referring to very specific instructional features to measure EL or GL, there may be a conceptual problem, because the expertise reversal effect shows that a particular instructional feature may be associated with GL (i.e., enhancing learning outcomes) for one learner and with EL (i.e., hindering learning outcomes) for another learner (Kalyuga et al., 2003). An alternative approach to the formulation of questions for EL and GL might solve this problem. Furthermore, the measurement could become more precise when using multiple items for each of the separate types of CL with a scale that is different from the scales used in previous research. It is not entirely clear to what extent workload and cognitive load refer to the same concept across settings, but the NASA-TLX is an example of an instrument that assesses work load on five 7-point scales. Increments of high, medium, and low estimates for each point result in 21 gradations on the scales (Hart & Staveland, 1988; Hilbert & Renkl, 2009; Zumbach & Mohraz, 2008).

A new instrument for the measurement of IL, EL, and GL

In this study, a new instrument for the measurement of IL, EL, and GL in complex knowledge domains was developed. The data for the present article were collected in four lectures and in a randomized experiment in statistics. Statistics is an important subject in many disciplines, jobs, study programs, and every-day situations. In this domain, abstract concepts are hierarchically organized and typically have little or no meaning outside the domain. Not only do learners need to learn formulas and how to apply them correctly, they also need to develop knowledge of key concepts and definitions, and have to learn to understand how statistical concepts are interrelated (Huberty, Dresden, & Bak, 1993). Although the latter requires intensive training, knowledge of key concepts and definitions and proficiency with basic formulas can be developed at an early stage (Leppink, Broers, Imbos, Van der Vleuten, & Berger, 2011, 2012a, b). Therefore, asking learners to rate difficulty or complexity of formulas, concepts, and definitions may be feasible at an early stage, whereas asking them to rate complexity of relationships between various concepts may not, because they may not yet be able to perceive any of these

relationships. With this in mind, the items displayed in Appendix 1 were developed.

Items 2 and 9 refer to formulas, whereas Items 1, 3, 7, 8, and 10 refer to concepts, definitions, or just the topics covered. Although Item 8 directly refers to understanding of statistics, of course the term "statistics" can be replaced by the term representing another complex knowledge domain if data are to be collected in, for example, mathematics, programming, physics, economics, or biology.

The ten items had been subjected to an online pilot study at a Belgian university (teaching in Dutch), involving 100 first year bachelor students in psychology, and 67 master students in psychology.

The present set of studies

In a set of four studies, all carried out in the same Dutch university, the performance of the new instrument was examined. In a first study (henceforth, Study I), the instrument was administered in a lecture in statistics for 56 PhD students in psychology and health sciences, and Hypotheses 1–3 were tested using principal component analysis:

> *Hypothesis 1*. Items 1, 2, and 3 all deal with complexity of the subject matter itself and are therefore expected to load on the factor of IL;
> *Hypothesis 2*. Items 4, 5, and 6 all deal with negative characteristics of instructions and explanations and are therefore expected to load on the factor of EL;
> *Hypothesis 3*. Items 7, 8, 9, and 10 all deal with the extent to which instructions and explanations contribute to learning and are therefore expected to load on the factor of GL.

In a second study (henceforth, Study II), we administered a questionnaire comprising these ten items and the aforementioned scales by Paas (1992) for CL, Ayres (2006) for IL, Cierniak et al. (2009) for EL, and Salomon (1984) for GL in a lecture in statistics for 171 second-year bachelor students in psychology, to test the first three and the following four hypotheses (i.e., Hypotheses 1–7) using confirmatory factor analysis:

> *Hypothesis 4*. Ayres's (2006) scale for IL loads on IL but *not* on EL or GL;
> *Hypothesis 5*. Cierniak et al.'s (2009) scale for EL loads on EL but *not* on IL or GL;
> *Hypothesis 6*. Salomon's (1984) scale for GL loads on GL but *not* on IL or EL;
> *Hypothesis 7*. Paas's (1992) scale for CL loads on IL, EL, and GL.

Hypotheses 4–7 received no support from the data in Study II. Ayres's (2006) scale for IL had a lower loading on IL than Items 1, 2, and 3, and it had a significant cross-loading on EL. Cierniak et al.'s (2009) scale for EL and Salomon's (1984) scale for GL diverged from the other items in the instrument, and Paas's (1992) scale for CL has relatively weak loadings on all three factors. Therefore, only Hypotheses 1–3 were tested using confirmatory factor analysis in a third study (henceforth, Study III). The data for this analysis were collected in a lecture in statistics for 136 third-year bachelor students in psychology, and in a lecture in statistics for 148 first-year bachelor students in health sciences. As Studies I, II, and III together provided support for Hypotheses 1–3, a three-factor approach for IL, EL, and GL was adopted in a fourth study (henceforth: Study IV).

In Study IV, a randomized experiment was conducted to examine the effects of experimental treatment and prior knowledge on CL, IL, EL, and GL, and learning outcomes. In this experiment, a total of 58 novice learners studied a problem either in a familiar format (textual explanation) and subsequently in an unfamiliar format (formula; $n = 29$) or in an unfamiliar format (formula) and subsequently in a familiar format (textual explanation; $n = 29$). Studies by Reisslein, Atkinson, Seeling, and Reisslein (2006) and Van Gog, Kester, and Paas (2011) have demonstrated that example-problem pairs are more effective for novices' learning than problem–example pairs. Even though both conditions receive the same tasks, the order matters, presumably because studying an example first induces lower EL and higher GL, allowing for schema building. That schema can subsequently be used when solving the problem. When solving a problem first, there is very high EL and little learning. In line with these findings, we expected that learners who studied the problem in a familiar (textual) format first would demonstrate better learning outcomes (because they could use what they had learned from the text to understand the formula) and respond with lower levels of EL and higher levels of GL. Further, we expected learners with more prior knowledge to demonstrate better learning outcomes and respond with lower levels of IL than less knowledgeable learners. Thus, Hypotheses 8–12 were tested in a randomized experiment:

> *Hypothesis 8*. Learners who have more prior knowledge experience lower IL than learners who have less prior knowledge;
> *Hypothesis 9*. Learners who have more prior knowledge demonstrate better learning outcomes than learners who have less prior knowledge;
> *Hypothesis 10*. Studying a problem first in a familiar format and subsequently in an unfamiliar format enhances learning outcomes more than studying the same problem first in an unfamiliar format and subsequently in a familiar format;

Hypothesis 11. Studying a problem first in a familiar format and subsequently in an unfamiliar format imposes less EL on a learner than studying the same problem first in an unfamiliar format and subsequently in a familiar format;

Hypothesis 12. Studying a problem first in a familiar format and subsequently in an unfamiliar format imposes more GL on a learner than studying the same problem first in an unfamiliar format and subsequently in a familiar format.

In the following discussion, methods and results are discussed for each of the studies separately. Next, findings and limitations are discussed for each of the studies, and implications for future research are discussed.

Study I: Exploratory analysis

Method

A total of 56 PhD students in the social and health sciences, who attended a lecture on multiple linear regression analysis and analysis of variance, completed the questionnaire. To avoid potential confounding from specific item-order effects, the items presented in Appendix 1 were counterbalanced in three orders: order A (n = 19), Items 1, 7, 4, 2, 8, 5, 3, 9, 6, and 10; order B (n = 20), Items 6, 10, 9, 3, 5, 8, 2, 7, 1, and 4; and order C (n = 17), Items 9, 3, 6, 8, 2, 4, 10, 5, 7, and 1. The forms were put in randomized order, so that people sitting next to each other were not necessarily responding to the same item at the same time. Although it was also part of the written instruction on the questionnaire that students received, 2 min of oral instruction was provided at the beginning of the lecture to emphasize that each of the items in the questionnaire referred to the lecture that students were going to attend. All students completed the questionnaire on paper at the very end of the lecture and returned it right away. The lecture lasted 120 min and students had a break of about 15 min somewhere halfway. This procedure was the same in the lectures in Study II and III.

Hypotheses 1–3 were tested using principal component analysis. Principal component analysis is a type of exploratory factor analysis, in that loadings from all items on all components are explored.

Results

Although the sample size of this lecture was rather small for a ten-item instrument, the distributional properties of the data allowed for this type of factor analysis [no outliers or extreme skewness or kurtosis, as well as sufficient interitem correlation; KMO = .692, Bartlett's $\chi^2(45) = 228$, $p < .001$]. In case of this type of small sample, principal component analysis is preferred to principal factor analysis because it is less dependent on assumptions (e.g., normally distributed residuals are assumed in the latter).

Oblique (i.e., Oblimin) rotation was performed to take the correlational nature of the components into account (orthogonal rotation assumes that the factors are uncorrelated). If the components underlying the ten items are as hypothesized—IL, EL, and GL—correlation between components is to be expected. For the knowledgeable learner, IL may be low and the instructional features that contribute to EL and GL, respectively may be different from the instructional features that contribute to EL and GL for less knowledgeable learners. Learners who experience extremely high IL and/or high EL may not be able or willing to engage in GL activities. Using oblique rotation in principal component analysis, the correlation between each pair of components is estimated and taken into account in the components solution. Means (and standard deviations, *SD*), skewness, kurtosis, and component loadings are presented in Table 1. No outliers were detected.

Figure 1 shows a component loading plot. The component loadings are in line with Hypotheses 1–3, and no cross-loadings above .40 are present. Although the absence of cross-loadings above .40 is a positive sign, given the limited sample size of n = 56, the component loadings reported in Table 1 only provide a preliminary indication of what the component solution may be. In Table 2, we present the correlations between the three components.

Reliability analysis for the three components revealed Cronbach's alpha values of .81 for Items 1, 2, and 3 (expected to measure IL); .75 for Items 4, 5, and 6 (expected to measure EL); and .82 for Items 7, 8, 9, and 10 (expected to measure GL).

Study II: Confirmatory analysis

Method

Data were collected in a lecture for 171 second-year bachelor students in psychology on one-way and two-way analysis of variance. We justified a different cohort of students for this second study, because both lectures covered topics at a comparable level of difficulty. The students from both cohorts had limited knowledge of the topics covered, and therefore the lectures were of a rather introductory level. Furthermore, if a three-factor structure underlies the items in an instrument, one would expect that three-factor structure to hold across cohorts and potentially across settings.

To test Hypotheses 4–7, we added four items to the ten items presented in Appendix 1 that were introduced previously in this article: Paas's (1992) scale, which is assumed

Table 1 Means (and *SD*), skewness, kurtosis, and component loadings in Study I

Component/Item	Mean (*SD*)	Skewness	Kurtosis	Component Loading		
				C1	C2	C3
First Component						
Item 7	7.21 (1.19)	–0.77	0.36	.92	.01	.08
Item 8	7.04 (1.68)	–1.65	4.73	.84	.01	.01
Item 9	6.82 (1.42)	–0.03	–0.49	.83	–.02	.01
Item 10	6.84 (1.56)	–0.98	1.82	.65	.02	–.08
Second Component						
Item 1	5.54 (2.03)	–0.73	0.06	–.07	.76	.12
Item 2	5.41 (2.47)	–0.55	–0.93	.05	.84	.06
Item 3	5.75 (2.23)	–0.59	–0.21	.05	.94	–.15
Third Component						
Item 4	1.89 (1.36)	0.38	–0.47	.03	–.05	.91
Item 5	1.73 (1.26)	–0.04	–1.02	.04	–.03	.88
Item 6	1.88 (1.44)	1.02	2.28	–.11	.14	.63

to be an estimator of CL; a nine-point version of Ayres's (2006) six-point rating scale for IL; a nine-point version of Cierniak et al.'s (2009) seven-point rating scale for EL; and a nine-point version of the seven-point rating scale for GL used by Cierniak et al., who adopted it from Salomon (1984). These four items, presented in Appendix 2, formed the first four items of the questionnaire.

The item order for the ten new items was the same as order C in Study I. The reason that nine-point scales were used for each of these four items is to ease the standardization and interpretation of outcomes in the confirmatory factor analysis. If these items measure what they have been expected to measure, using a nine-point scale should cause no harm to the measurement. For example, higher EL should still be reflected in higher ratings on the nine-point version of Cierniak et al.'s (2009) seven-point rating scale for EL.

As in the principal component analysis on the data obtained in Study I, in the confirmatory factor analysis on the data in Study II, the correlation between each pair of factors was estimated and taken into account in the factor solution.

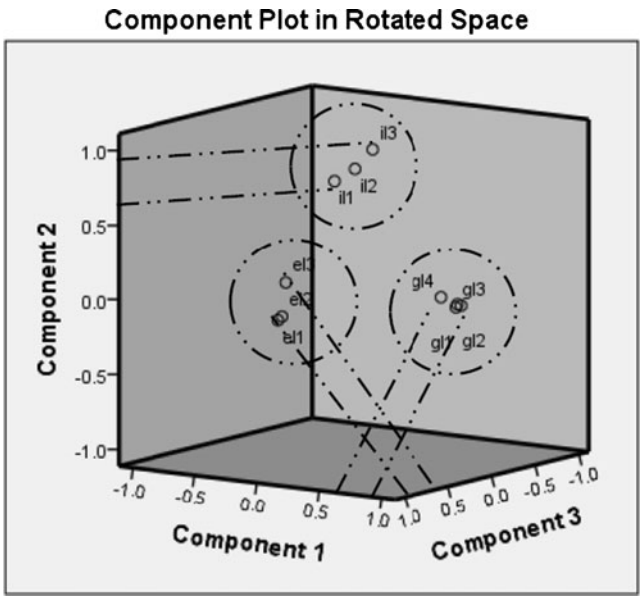

Fig. 1 Component loading plot for Study I

Results

In Table 3, we present means (and *SD*), skewness, and kurtosis, as well as squared multiple correlations (R^2) of each of the items administered in Study II. The R^2 is an indicator of item reliability and should preferably be .25 or higher.

The R^2 values reported in Table 3 and the factor loadings presented in Table 4 indicate that Cierniak et al.'s (2009) scales for EL and GL diverge from the other items in the instrument.

In addition, Paas's (1992) scale for CL has relatively weak loadings on all three factors, maybe due to capturing *overall* load, whereas all other items in the questionnaire focus on a *specific* type of load. Although the loading of .61 for Ayres's (2006) scale for IL could be acceptable from the loading point of view, the

Table 2 Component correlations in Study I

Component Pair	Correlation
Component 1–Component 2	.05
Component 1–Component 3	–.31
Component 2–Component 3	.27

Table 3 Means (and *SD*), skewness, and kurtosis in Study II

Factor/Item	Mean (*SD*)	Skewness	Kurtosis	R^2
Nine-Point Versions of Existing Scales (1–9)				
Paas	5.64 (1.40)	–0.57	0.09	.25
Ayres	5.15 (1.37)	–0.24	1.15	.38
Cierniak et al.	4.35 (1.36)	0.29	0.42	.18
Salomon	6.02 (1.66)	–0.64	–0.07	.23
New Items (0–10)				
Item 1	4.94 (2.06)	–0.26	–0.26	.50
Item 2	5.08 (2.21)	–0.28	–0.47	.82
Item 3	5.11 (2.19)	–0.28	–0.60	.71
Item 4	2.13 (1.90)	1.17	1.44	.86
Item 5	2.16 (1.59)	0.61	0.17	.56
Item 6	2.56 (2.23)	1.09	0.97	.43
Item 7	6.60 (1.65)	–0.85	1.77	.68
Item 8	6.37 (1.63)	–0.80	0.95	.76
Item 9	6.57 (1.68)	–1.01	1.24	.60
Item 10	6.30 (1.67)	–1.11	2.04	.63

modification indices reveal a significant cross-loading on EL, indicating that it may diverge from the other items that are expected to measure IL. In line with this, both its factor loading and its R^2 are lower than the factor loadings and R^2 of the other items that load on IL and have no significant cross-loadings.

Table 4 Factor loadings for each of the 14 items administered in Study II

Factor/Item	Factor Loading	*SE*	*t* Value	*p* Value
First Factor: IL				
Paas	.26	.097	2.70	.007
Ayres	.62	.053	11.67	<.001
Item 1	.71	.044	16.17	<.001
Item 2	.90	.024	36.94	<.001
Item 3	.84	.029	28.92	<.001
Second Factor: EL				
Paas	.00	.094	0.02	.99
Cierniak et al.	.42	.069	6.10	<.001
Item 4	.93	.031	29.75	<.001
Item 5	.75	.040	18.60	<.001
Item 6	.66	.050	13.25	<.001
Third Factor: GL				
Paas	.35	.083	4.25	<.001
Salomon	.48	.063	7.62	<.001
Item 7	.83	.031	26.88	<.001
Item 8	.87	.026	33.72	<.001
Item 9	.77	.037	21.04	<.001
Item 10	.79	.034	23.14	<.001

In the present study design, we cannot answer the question why these measures diverge, or which of the measures is a better measure of the different types of load, because the instructional tasks used in our study varied extensively from the prior studies. However, given that the ten recently developed items appear to form a three-factor solution from which the other four items diverge from, we continued by testing a model with only the ten recently developed items. The three factors are significantly correlated: the correlation between IL and EL is .41 ($p < .001$), the correlation between IL and GL is .33 ($p < .001$), and the correlation between EL and GL is –.19 ($p = .025$). Two additional residual covariance paths were included to the model—namely, between Item 7 and Item 9 and between Item 9 and Item 10. Item 9 asks students to rate the extent to which the activity contributed to their understanding of formulas, whereas Items 7 and 10 refer more to verbal information. These residual covariance paths were included, because the three lecturers involved in Study II and Study III were different in terms of emphasis on verbal explanation versus formulaic explanation.

Table 5 contains factor loadings of Items 1–10 in Study II and the correlations of the two residual covariance paths. The two residual covariance paths have small coefficients, and one of them was not statistically significant. We find $\chi^2(30) = 62.36$, $p < .001$, CFI = .965, TLI = .947, RMSEA = .079. The modification indices do not provide any meaningful suggestions for additional paths. Although the CFI and TLI appear to indicate that we have a good fitting model, the RMSEA is on the edge (i.e., above .08 is inadequate, values around .06 are

Table 5 Factor loadings for each of the ten recently developed items administered in Study II

Factor/Item	Factor Loading	*SE*	*t* Value	*p* Value
First Factor: IL				
Item 1	.68	.046	14.83	<.001
Item 2	.93	.027	34.40	<.001
Item 3	.84	.032	26.07	<.001
First Factor: EL				
Item 4	.95	.034	27.79	<.001
Item 5	.74	.042	17.54	<.001
Item 6	.65	.051	12.72	<.001
First Factor: GL				
Item 7	.79	.036	21.62	<.001
Item 8	.91	.028	32.53	<.001
Item 9	.73	.046	15.84	<.001
Item 10	.80	.035	22.69	<.001
Residual Covariance				
Item 7, Item 9	.29[a]	.090	3.19	<.001
Item 9, Item 10	–.03[a]	.10	–0.35	.73

[a] This is a correlation, not a factor loading.

acceptable, and values of .05 and lower are preferred). We decided to test this model on the new data collected in two lectures in Study III.

Study III: Cross-validation

Method

The instrument was administered in a lecture for 136 third-year bachelor students in psychology on logistic regression and in a lecture for 148 first-year bachelor students in health sciences on null hypothesis significance testing. In the lecture on logistic regression, the items were asked in the order presented in Appendix 1. In the lecture on null hypothesis significance testing, the items presented in Appendix 1 were presented in three orders: the order in Appendix 1 ($n = 50$), as well as order D ($n = 49$)—Items 1, 5, 10, 2, 6, 3, 7, 8, 4, and 9—and order E ($n = 49$)—Items 5, 9, 1, 3, 10, 4, 6, 8, 2, and 7 (i.e., orders D and E were used because the orders were different from orders A, B, and C used previously). The forms were put in randomized order, so that people sitting next to each other were not necessarily answering the same questions.

We are aware that the cohorts in Study III differ from each other in terms of knowledge of statistics and that both cohorts differ from the cohorts in Study I and Study II. All four lectures in the three studies, however, covered content that had not been taught to these cohorts before and were therefore of a rather introductory level. Furthermore, administering an instrument in different cohorts potentially increases variability of responses and enables the stability of a factor solution. If a factor solution is consistent across datasets, this is an indicator of the stability of the solution.

Results

Table 6 shows the factor loadings of the ten items and the correlations of the two residual covariance paths in the lecture on logistic regression.

The residual covariance that had been statistically significant in Study II was not statistically significant in the lecture on logistic regression, whereas the other residual covariance had a moderate coefficient and was statistically significant.

The three factors were significantly correlated: The correlation between IL and EL was .61 ($p < .001$), the correlation between IL and GL was –.36 ($p < .001$), and the correlation between EL and GL was –.56 ($p < .001$). The analysis yielded $\chi^2(30) = 35.036$, $p = .24$, CFI = .995, TLI = .992, RMSEA = .035. Table 7 contains the factor loadings of the ten items and the correlations of the two residual covariance paths in the lecture on null hypothesis significance testing.

Table 6 Factor loadings for each of the ten recently developed items administered in the lecture on logistic regression

Factor/Item	Factor Loading	*SE*	*t* Value	*p* Value
First Factor: IL				
Item 1	.82	.035	23.27	<.001
Item 2	.81	.035	23.17	<.001
Item 3	.92	.026	35.74	<.001
First Factor: EL				
Item 4	.83	.044	18.95	<.001
Item 5	.69	.056	12.43	<.001
Item 6	.77	.049	15.88	<.001
First Factor: GL				
Item 7	.86	.027	31.17	<.001
Item 8	.99	.017	57.82	<.001
Item 9	.78	.035	22.16	<.001
Item 10	.79	.035	22.90	<.001
Residual Covariance				
Item 7, Item 9	.10[a]	.083	1.15	.25
Item 9, Item 10	.43[a]	.075	5.74	<.001

[a] This is a correlation, not a factor loading.

Both residual covariance paths were close to zero and not statistically significant in the lecture on null hypothesis significance testing. Furthermore, only IL and EL were significantly correlated: The correlation between IL and EL was .25 ($p = .007$), the correlation between IL and GL

Table 7 Factor loadings for each of the ten recently developed items administered in the lecture on null hypothesis significance testing

Factor/Item	Factor Loading	*SE*	*t* Value	*p* Value
First Factor: IL				
Item 1	.71	.052	13.63	<.001
Item 2	.83	.046	18.09	<.001
Item 3	.78	.048	16.26	<.001
First Factor: EL				
Item 4	.88	.038	23.14	<.001
Item 5	.76	.045	17.10	<.001
Item 6	.78	.044	17.74	<.001
First Factor: GL				
Item 7	.89	.026	33.95	<.001
Item 8	.89	.026	34.07	<.001
Item 9	.76	.047	15.99	<.001
Item 10	.82	.032	25.39	<.001
Residual Covariance				
Item 7, Item 9	.03[a]	.149	0.18	.86
Item 9, Item 10	–.06[a]	.119	–0.49	.63

[a] This is a correlation, not a factor loading.

Table 8 R^2 values for each of the ten items in the final model, along with Cronbach's alpha values, per scale in Study II and Study III

	R^2 Values of Item and Cronbach's Alphas of Scales		
		Study III	
Scale / Item	Study II	Logistic Regression	Hypothesis Testing
IL	.85[a]	.88[a]	.81[a]
Item 1	.46	.68	.51
Item 2	.86	.66	.69
Item 3	.70	.85	.61
EL	.80[a]	.81[a]	.85[a]
Item 4	.90	.69	.78
Item 5	.55	.48	.58
Item 6	.42	.60	.61
GL	.89[a]	.93[a]	.91[a]
Item 7	.62	.73	.80
Item 8	.82	.99	.80
Item 9	.53	.61	.58
Item 10	.64	.63	.68

[a] These are Cronbach's alpha values.

was .04 (p = .65), and the correlation between EL and GL was –.11 (p = .24). These results yielded $\chi^2(30) = 30.298$, p = .45, CFI = 1.000, TLI = .999, RMSEA = .008. Table 8 shows the R^2 values for each of the ten items in the final model and Cronbach's alpha values per scale for the lectures in Studies II and III.

The lowest R^2 value was .42 in Study II (Item 6, which appears to be an indicator of EL), which indicates that every item has a sufficient amount of variance in common with other items in the questionnaire.

Study IV: Experiment

Method

A total of 58 university freshmen who were about to enter a course in basic inferential statistics participated in a randomized experiment, in which two groups studied a problem on conditional and joint probabilities in counterbalanced order. Prior knowledge of conditional and joint probabilities was assessed prior to the study, and immediately after the study a posttest on conditional and joint probabilities was administered.

The students had a stake in the experiment, as the content of the experiment would form the content of the first week in their upcoming statistics course. The students were informed that they would participate in a short experiment and that this experiment would be followed by a one-hour lecture in which the content covered in the experiment—conditional and joint probabilities—would be explained. Participation in the experiment lasted 45 min, and the subsequent lecture lasted 60 min.

In the lecture, conditional and joint probabilities as well as frequent misconceptions on these topics were discussed by a statistics teacher. The lecture was interactive; not only did the lecturer explain the concepts of conditional and joint probability, the lecturer also stimulated students in the audience who knew the answer to the problem presented on the screen to explain their reasoning to their peers. After the lecture, students were also debriefed about the setup of the experiment. Finally, lecture slides as well as correct calculations and answers to all the items in the prior knowledge test and posttest were provided to the students, and students were allowed to stay in touch via email with the lecturer to ask questions on the content or on the provided materials.

From an ethical perspective, we wanted to avoid potential disadvantage for individual students due to them having participated in a specific treatment order condition. Through an additional lecture for all participating students together, we expected to compensate for unequal learning outcomes resulting from the experiment. From a motivational perspective, we expected that providing students with feedback on their performance in (as well as after) such a lecture would stimulate students to take the experiment seriously, which could reduce noise in their responses to the various items.

At the very start of the meeting, all students completed the prior knowledge test on conditional and joint probabilities that is presented in Appendix 3.

To reduce guessing behavior, multiple choice items were avoided and open-answer questions were used. Students had to calculate a conditional probability in the first question and a joint probability in the second question. As expected, both questions were of a sufficient difficulty level in that they did not lead to extremely low correct response proportions: the first question yielded fifteen correct responses (about 26 % of the sample) and the second question yielded 31 correct responses (about 53 % of the sample). At the end of the prior knowledge test, students completed the same questionnaire as presented in Appendix 1.

Next, students were assigned randomly to either of two treatment order conditions. In both conditions, students were presented the same problem on conditional and joint probabilities in two modes: in an explanation of six lines text, and in formula notation. In treatment order condition TF, students first studied the text explanation (T) and then the formula explanation (F), and in condition FT, the order was the other way around. The two presentation formats—text and formula—are presented in Appendix 4.

Students reported, as expected, that they were not familiar with the specific notation of conditional probabilities like P(man | psychology). In both treatment conditions, students completed the same questionnaire as they completed after the prior knowledge test and after each study format. The

two formats were not presented simultaneously; students received the two formats in counterbalanced order, and which format they received first depended on the treatment order condition.

To assess learning outcomes, a five-item posttest on conditional and joint probabilities was administered. The items were similar to the questions in the prior knowledge test and resembled the problem studied in the two formats, only more difficult to avoid potential ceiling effects for some items. Correct response rate on an item varied from sixteen respondents (about 31 % of the sample) to 32 respondents (about 55 % of the sample). The average number of correctly responded items was 1.97, and Cronbach's alpha of the five-item scale was .79. Having completed the five-item posttest, students completed the same questionnaire as they completed after the prior knowledge test and after the two study formats. Thus, we had four measurements for all the CL-related items per participating student. Completed questionnaires were checked for missing responses right away, which confirmed that all participants responded to all the items in the questionnaire. Likewise, on the prior knowledge test and posttest, no missing responses were found.

Results

The reliability analysis revealed that Items 1, 2, and 3 form a homogeneous scale, and when we added Ayres's (2006) item for IL, the Cronbach's alpha of the scale remained more or less the same. Furthermore, Items 4, 5, and 6 form a scale for which Cronbach's alpha decreased considerably in three of the four measurements when Cierniak et al.'s (2009) item for EL was added. Similarly, Items 7, 8, 9, and 10 form a homogeneous scale for which Cronbach's alpha decreased considerably when Salomon's (1984) item for GL was added. Finally, Paas's (1992) item for CL appears to be correlated to the items that aim to measure IL only, and adding Paas's item to the scale with Items 1, 2, 3, and Ayres's item for IL did not lead to remarkable changes in Cronbach's alpha. These findings are presented in Table 9 for the four time points (i.e., after prior knowledge test, after text format, after formula format, after posttest), respectively.

Table 10 shows the means and standard deviations for each of the three scales of Items 1–10 and for the four 9-point scales at each of the four time points, per treatment order condition (i.e., TF and FT).

The somewhat lower Cronbach's alpha value for the scale of Items 4, 5, and 6 after the prior knowledge test and after the posttest may be a consequence of restriction of range effects. After both treatment formats, there is more variation in scores on this scale and Cronbach's alpha values of the scale are within the expected range. As expected, the average score on this scale was highest after the formula format in treatment condition FT, where students were confronted with the formula format before they received the text format.

Linear contrast analysis for the effect of prior knowledge (number of items correct: 0, 1, or 2) on posttest performance (0–5) revealed a linear effect, $F(1, 24) = 8.973$, $p < .01$, $\eta^2 = .134$, and the deviation was not statistically significant, $F(1, 7) = 2.76$, $p = .10$, $\eta^2 = .041$. We therefore included prior knowledge as a linear predictor in our subsequent regression analysis for posttest performance. None of the CL-related scores obtained after the prior knowledge test, after the text format, and after the formula format contributed significantly to posttest performance. In Table 11, we present the results of an analysis of covariance (ANCOVA) model for posttest performance using prior knowledge score, treatment order, and the average on the scale of Items 7, 8, 9, 10—the four items that are supposed to measure GL—as predictors after the posttest. Of the other CL-related scales after the posttest, none contributed significantly to posttest performance, which makes sense because only GL activities should contribute to learning and result in better learning outcomes.

In line with Hypothesis 9, a higher prior knowledge score was a statistically significant predictor for higher posttest performance. Furthermore, posttest performance was non-significantly worse in the TF condition, meaning we have

Table 9 Cronbach's alphas of three scales in Study IV

Scale	Prior	Text	Formula	Posttest
Items 1, 2, 3	.86	.87	.91	.89
Items 1, 2, 3 + Ayres (2006)	.86	.89	.89	.89
Items 1, 2, 3 + Ayres + Paas (1992)	.86	.89	.89	.89
Items 4, 5, 6	.71	.85	.87	.63
Items 4, 5, 6 + Cierniak et al. (2009)	.54	.80	.82	.67
Items 4, 5, 6 + Cierniak et al. + Paas	.50	.76	.78	.64
Items 7, 8, 9, 10	.94	.97	.94	.96
Items 7, 8, 9, 10 + Salomon (1984)	.83	.89	.85	.87
Items 7, 8, 9, 10 + Salomon + Paas	.74	.84	.80	.81

Table 10 Mean (and *SD*) for each of the three scales of Items 1–10 and for the four 9-point scales, per treatment order condition in Study IV

Scale/Item	Text Formula (TF)	Formula Text (FT)
After Prior Knowledge		
Items 1, 2, 3	3.17 (2.21)	4.59 (1.96)
Items 4, 5, 6	1.56 (1.47)	2.06 (1.80)
Items 7, 8, 9, 10	3.49 (2.33)	3.54 (1.84)
Paas (1992)	5.52 (1.55)	5.52 (1.41)
Ayres (2006)	5.48 (2.03)	6.10 (1.47)
Cierniak et al. (2009)	5.62 (1.80)	5.31 (1.63)
Salomon (1984)	6.07 (1.79)	6.34 (1.47)
After Text Format		
Items 1, 2, 3	5.05 (2.48)	4.48 (2.14)
Items 4, 5, 6	3.26 (2.07)	3.54 (2.50)
Items 7, 8, 9, 10	3.52 (2.58)	4.83 (1.77)
Paas	6.31 (1.54)	5.76 (1.33)
Ayres	6.28 (1.75)	5.72 (1.41)
Cierniak et al.	6.10 (1.63)	5.34 (1.45)
Salomon	6.76 (1.38)	6.48 (1.41)
After Formula Format		
Items 1, 2, 3	4.31 (2.41)	5.09 (1.75)
Items 4, 5, 6	2.24 (2.21)	4.68 (2.40)
Items 7, 8, 9, 10	4.46 (2.38)	4.31 (1.61)
Paas	5.59 (1.76)	5.83 (1.26)
Ayres	5.59 (1.57)	5.76 (1.19)
Cierniak et al.	5.14 (1.58)	5.62 (1.43)
Salomon	6.07 (1.60)	6.21 (1.40)
After Posttest		
Items 1, 2, 3	4.97 (2.28)	5.22 (2.13)
Items 4, 5, 6	2.14 (1.32)	2.41 (1.91)
Items 7, 8, 9, 10	4.40 (2.31)	4.71 (1.66)
Paas	6.76 (1.30)	6.66 (1.14)
Ayres	6.38 (1.66)	6.52 (1.18)
Cierniak et al.	6.03 (1.52)	5.76 (1.46)
Salomon	7.28 (1.33)	7.00 (1.23)

no support for Hypothesis 10. Finally, there is limited evidence that higher scores on the scale of Items 7, 8, 9, and 10—intended to measure GL—predict higher posttest performance (η^2 = .064). It is possible that students were still learning to a more or lesser extent while completing the posttest.

For the effects of prior knowledge and experimental treatment on IL, EL, and GL, as measured by the scales of Items 1–10, mixed linear models with Toeplitz as covariance structure provided the best solution for analysis.

Table 12 shows the outcomes of this model for average IL (i.e., Items 1, 2, and 3). In line with Hypothesis 8, the model presented in Table 12 indicates that more prior knowledge predicts lower IL. Furthermore, presenting the formula format before the text format appears to lower IL experienced when studying the text presentation but not when studying the formula presentation.

Table 13 presents the outcomes of the model for average EL (i.e., Items 4, 5, and 6). Confirming Hypothesis 11, the model presented in Table 13 indicates that when the formula format is presented before the text format, EL is elevated significantly for the formula format.

Finally, Table 14 reveals the outcomes of the model for average GL (i.e., Items 7, 8, 9, and 10). The model presented in Table 14 indicates that the text format imposes significantly more GL when presented after the formula format. On the one hand, one may argue that the formula format confronted students with difficulties, leading them to invest more GL activities when the textual explanation was provided. On the other hand, however, no significantly elevated posttest performance was detected.

Discussion

In this section, findings and limitations are discussed for the four studies, and implications for future research are discussed.

Exploratory analysis

Although the sample size was small for a ten-item instrument, the principal component analysis in Study I provided preliminary support for Hypotheses 1, 2, and 3. Also, as one would expect, the components that are expected to be EL and GL are negatively correlated. Furthermore, the components that are expected to measure IL and GL have a correlation around zero. The relationship between IL and GL may not be linear. Extremely low as well as extremely high levels of IL may lead to limited GL activity. On the one hand, if a learning task is too easy for a student, the

Table 11 ANCOVA model for posttest performance, using as covariates prior knowledge score, treatment order, and average score on the scale of Items 7–10 after the posttest in Study IV

Effect	Coefficient	Standard Error	*t*(55)	*p* Value
Intercept	0.63	0.59	1.06	.29
Prior knowledge score	0.96	0.33	2.87	<.01
Order	–0.61	0.42	–1.44	.15
Average, Items 7–10	0.19	0.11	1.80	.08

Order coding: 0 = TF, 1 = FT.

Table 12 Mixed linear model for IL in Study IV

Effect	Coefficient	Standard Error	*t* Value	*p* Value
Intercept	4.63	0.40	10.96	<.01
Prior knowledge score	–1.20	0.29	–4.09	<.01
Order	0.90	0.39	2.30	.03
Text (dummy)	1.57	0.40	3.90	<.01
Formula (dummy)	0.82	0.31	2.64	<.01
Posttest (dummy)	1.21	0.34	3.59	<.01
Order × Text	–1.37	0.50	–2.72	<.01

Order coding: 0 = TF, 1 = FT.

explanations and instructions in the task may not contribute to actual learning on the part of that student. On the other hand, if a learning task is too complex for a particular student, cognitive capacity for GL activity may be very limited. Finally, the components that are expected to measure IL and EL have a moderately positive correlation.

Confirmatory support for a three-factor model

The fact that the items presented in Appendix 1 have different factor loadings than the previously developed scales for measuring the different types separately is interesting, but also hard to explain on the basis of the present data. Moreover, since no learning outcomes were measured after the lectures, these studies do not provide insight in how the various scales are related to learning outcomes. For this reason, we conducted the randomized experiment in Study IV (1) to examine how different scales vary in two different experimental conditions that we expected to lead to differential effects on IL, EL, and GL, and (2) to examine how the various scales are related to learning outcomes. Together, the results of Study II and Study III provide support for the three-component solution found in Study I.

The high item reliabilities (i.e., R^2 values), high Cronbach's alpha values, and high fit indices (i.e., CFI and TLI) across lectures in studies I to III, and the low RMSEA in two of the three confirmatory factor analyses support our expectation that a three-factorial structure underlies Items 1–10. It has been suggested that the concept of GL should be redefined as referring to actual working memory resources devoted to dealing with IL rather than EL (Kalyuga, 2011; Sweller, 2010). Kalyuga suggested that "the dual intrinsic/extraneous framework is sufficient and non-redundant and makes boundaries of the theory transparent" (2011, p. 1). Contrary to EL and IL, GL "was added to the cognitive framework based on theoretical considerations rather than on specific empirical results that could not be explained without this concept" (Kalyuga, 2011, p. 1). The present findings suggest, however, that such a two-factor framework may not be sufficient; the three-factor solution is consistent across lectures.

On the use of different cohorts in Studies I, II, and III

We justified the use of different cohorts of students in the four lectures studied. If a factor solution is consistent across these varied datasets, this is an indicator of the stability of the solution. The reason that we chose two lectures instead of one lecture in Study III was to have two independent lectures additional to the lecture Study II to test the hypothesized three-factor model. However, the use of different cohorts and different lecturers may introduce confounds, which may partly explain why the correlation between factor pairs and the residual covariances are somewhat different correlations across lectures.

Cohort-related factors may form one source of confounding. PhD students—and to some extent also advanced bachelor students—are, more than university freshmen, aware of the importance of statistics in their later work.

Table 13 Mixed linear model for EL in Study IV

Effect	Coefficient	Standard Error	*t* Value	*p* Value
Intercept	2.07	0.41	5.07	<.01
Prior knowledge score	–0.57	0.30	–1.87	.07
Order	0.39	0.41	0.93	.36
Text (dummy)	1.59	0.31	5.18	<.01
Formula (dummy)	0.61	0.39	1.58	.12
Posttest (dummy)	0.47	0.25	1.88	.07
Order × Formula	2.07	0.51	4.09	<.01

Order coding: 0 = TF, 1 = FT.

Table 14 Mixed linear model for GL in Study IV

Effect	Coefficient	Standard Error	*t* Value	*p* Value
Intercept	3.33	0.42	7.87	<.01
Prior knowledge score	0.20	0.32	0.62	.54
Order	0.07	0.42	0.16	.87
Text (dummy)	0.03	0.38	0.09	.93
Formula (dummy)	0.87	0.30	2.87	<.01
Posttest (dummy)	1.04	0.32	3.28	<.01
Order × Text	1.25	0.47	2.63	.01

Order coding: 0 = TF, 1 = FT.

Teaching style may form a second source of confounding: Whereas some lecturers emphasize conceptual understanding, others emphasize formulas and computations. In a lecture in which the focus is on conceptual understanding rather than on formulas, Item 9 may be a somewhat weaker indicator of GL. If the focus in a lecture is on formulas and conceptual understanding is of minor importance, Item 10 may be a somewhat weaker indicator of GL.

A third potential source of confounding in these studies was the subject matter. Whereas the lectures in Study I and Study II covered similar topics, the lectures in Study III were on different topics, which could have affected the measurement of the different types of load.

Future validation studies should administer this instrument in different lectures of a number of courses given by the same lecturers and for the same cohorts of students, repeatedly, to estimate the magnitude of student-related, teacher-related, and subject-related factors in item response and to examine the stability of the three-factor model across time.

Additional support for the three-factor solution in the experiment

The experiment in Study IV provides evidence for the validity of the three-factor solution underlying Items 1–10. First of all, as expected, higher prior knowledge predicted lower IL throughout the study (all four time points) and higher posttest performance. More knowledgeable learners have more elaborated knowledge structures in their long-term memory and are therefore expected to experience lower IL due to novelty of elements and element interactivity in a task (Kalyuga, 2011; Van Merriënboer & Sweller, 2005).

Secondly, as expected, EL during learning was higher when a problem to be studied was presented first in a format learners were not familiar with (the formula format); however, learners appeared to engage more in GL activities if the problem was subsequently presented in a format they were familiar with (the text format). Also, the known format was reported to impose less IL when presented after the unknown format. Although the students who received the unknown (formula) format first complained that it was difficult and responded to the questionnaire with higher rates of EL after the unknown format, they subsequently responded with lower rates of IL and higher rates of GL after the text format. These findings are difficult to explain, and suggest that order effects may influence the IL that is experienced by a learner. A limitation of this study was that only one posttest was administered after studying both formats, so we cannot determine to what extent *each* of the formats separately contributed to posttest performance. Future studies should include a test after each format instead of only after both formats. This may also provide more insight into why, in the present experiment, no negative effects of EL on learning performance were found. It is possible that higher EL experienced among students who received the formula format first compensated by increased investment in GL activities in the subsequent study in the text format.

Finally, there is limited evidence that higher scores on GL after the posttest predict higher posttest performance. New experiments, using larger sample sizes, are needed to further investigate this finding.

Question wording effects

More experimentation is also needed to examine across a wide range of learning tasks and contexts the correlations between the items presented in Appendix 2 and the three factors that underlie Items 1–10. Specific wording effects may play a role. For example, Paas's (1992) item for CL directly asks how much effort learners *invest* in an activity. This "investment" term is not used in any of the other items included. In addition, the question "how difficult it is to learn with particular material" could refer to EL for some learners and to IL for other learners. New studies should examine qualitatively how exactly learners *interpret* these items across a range of tasks.

Implications and suggestions for future research

For the present set of studies, the statistics knowledge domain was chosen because this is a complex knowledge domain that is important in many professions and academic curricula, and potentially even in everyday contexts. As

with the items developed by Paas (1992), Ayres (2006), Cierniak et al. (2009), and Salomon (1984), however, the intended applicability of Items 1–10 is not restricted to a particular knowledge domain. With minor adjustments (e.g., "statistics" in some items), these items could be used in research in other complex knowledge domains.

Finally, studies combining the subjective measures presented in this article—including the four items developed by Paas (1992), Ayres (2006), Cierniak et al. (2009), and Salomon (1984)—and biological measures such as eye-tracking (Holmqvist et al., 2011; Van Gog & Scheiter, 2010) may lead to new insights on convergence between biological and subjective measures and on what these different types of measures are measuring. If both biological and subjective measures measure the same constructs—in this context, IL, EL, and GL, and potentially even overall CL as a function of these three types of CL—one would expect high and positive correlations between these measures across educational settings. If such correlations are found, that may imply for measurement that using either of two types is potentially sufficient in educational studies. If other types of correlations are found, this opens doors for new research on why and under what circumstances the different types of measures diverge.

Appendix 1: A ten-item questionnaire for the measurement of IL (Items 1, 2, and 3), EL (Items 4, 5, and 6), and GL (Items 7, 8, 9, and 10)

All of the following questions refer to the activity (lecture, class, discussion session, skills training or study session) that just finished. Please respond to each of the questions on the following scale (0 meaning *not at all the case* and 10 meaning *completely the case*).

0 1 2 3 4 5 6 7 8 9 10

[1] The topic/topics covered in the activity was/were very complex. (il1 in Fig. 1)
[2] The activity covered formulas that I perceived as very complex. (il2 in Fig. 1)
[3] The activity covered concepts and definitions that I perceived as very complex. (il3 in Fig. 1)
[4] The instructions and/or explanations during the activity were very unclear. (el1 in Fig. 1)
[5] The instructions and/or explanations were, in terms of learning, very ineffective. (el2 in Fig. 1)
[6] The instructions and/or explanations were full of unclear language. (el3 in Fig. 1)
[7] The activity really enhanced my understanding of the topic(s) covered. (gl1 in Fig. 1)
[8] The activity really enhanced my knowledge and understanding of statistics. (gl2 in Fig. 1)
[9] The activity really enhanced my understanding of the formulas covered. (gl3 in Fig. 1)
[10] The activity really enhanced my understanding of concepts and definitions. (gl4 in Fig. 1)

Appendix 2: Four additional items for data collection in Study II—Item 1, expected to measure CL (Paas, 1992); Item 2, expected to measure IL (Ayres, 2006); Item 3, expected to measure EL (Cierniak et al., 2009); and Item 4, expected to measure GL (Salomon, 1984)

[1] Please choose the category (1, 2, 3, 4, 5, 6, 7, 8, or 9) that applies to you: In the lecture that just finished I invested
1. very, very low mental effort / 2. very low mental effort / 3. low mental effort / 4. rather low mental effort / 5. neither low nor high mental effort / 6. rather high mental effort / 7. high mental effort/ 8. very high mental effort / 9. very, very high mental effort
[2] Please choose the category (1, 2, 3, 4, 5, 6, 7, 8, or 9) that applies to you: The lecture that just finished was
1. very, very easy / 2. very easy / 3. easy / 4. rather easy / 5. neither easy nor difficult / 6. rather difficult / 7. difficult / 8. very difficult / 9. very, very difficult
[3] Please choose the category (1, 2, 3, 4, 5, 6, 7, 8, or 9) that applies to you: To learn from the lecture was
1. very, very easy / 2. very easy / 3. easy / 4. rather easy / 5. neither easy nor difficult / 6. rather difficult / 7. difficult / 8. very difficult / 9. very, very difficult
[4] Please choose the category (1, 2, 3, 4, 5, 6, 7, 8, or 9) that applies to you: How much did you concentrate during the lecture?
1. very, very little / 2. very little / 3. little / 4. rather little / 5. neither little nor much / 6. rather much / 7. much / 8. very much / 9. very, very much

Appendix 3: Prior knowledge test in Study IV

Question 1

Student population *X*. consists of 600 men and 400 women. There are 200 chemistry students and of these 200 chemistry students, 100 are women. We now draw one student. What is the probability of a chemistry student, given that the student is a man?

Question 2

Student population *X*. consists of 600 men and 400 women. There are 300 business students and half of them are men. If we draw at random one student from student population *X*.,

what is the probability that the student happens to be a male business student?

Appendix 4: Presentation formats (text and formula) in Study IV

Text

If we draw at random 1 student from student population *X*, the probability that the student is a man is 0.5, and the probability that the student studies psychology is 0.2. The probability that the student is a man, given that the student studies psychology, is 0.3. From this follows that the probability that our student is a male psychology student is 0.2 times 0.3 and this is 0.06. The probability that our student studies psychology, given that the student is a man, can now be calculated by dividing the probability of a male psychology student by the probability that the student is a man, or: 0.06 / 0.5 = 0.12.

Formula

If we draw at random 1 student from student population *X*:

$$
\begin{aligned}
&P(\text{man}) = 0.5\\
&P(\text{psychology}) = 0.2\\
&P(\text{man}|\text{psychology}) = 0.3\\
&P(\text{man and psychology}) = P(\text{psychology})\\
&\qquad\qquad\qquad\qquad \times P(\text{man} \mid \text{psychology})\\
&\qquad\qquad\qquad = 0.2 \times 0.3 = 0.06\\
&P(\text{psychology} \mid \text{man}) = P(\text{man and psychology})/P(\text{man})\\
&\qquad\qquad\qquad = 0.06/0.5 = 0.12
\end{aligned}
$$

References

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.

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

Berthold, K., & Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. *Journal of Educational Psychology, 101,* 70–87.

Camp, G., Paas, F., Rikers, R. M. J. P., & Van Merriënboer, J. J. G. (2001). Dynamic problem selection in air traffic control training: A comparison between performance, mental effort, and mental efficiency. *Computers in Human Behavior, 17,* 575–595.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8,* 293–332.

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.

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology, 79,* 347–362.

DeLeeuw, 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. doi: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.

Fischer, F. (2002). Gemeinsame Wissenskonstruktion—Theoretische und methodologische Aspekte [Joint knowledge construction—Theoretical and methodological aspects]. *Psychologische Rundschau, 53,* 119–134.

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. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam: 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.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.

Huberty, C. J., Dresden, J., & Bak, B. (1993). Relations among dimensions of statistical knowledge. *Educational and Psychological Measurement, 53,* 523–532.

Kalyuga, S. (2009). Knowledge elaboration: A cognitive load perspective. *Learning and Instruction, 19,* 402–410.

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review, 23,* 1–19.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38,* 23–31.

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.

Kalyuga, S., & Hanham, J. (2011). Instructing in generalized knowledge structures to develop flexible problem solving skills. *Computers in Human Behavior, 27,* 63–68.

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.

Leppink, J., Broers, N. J., Imbos, T., Van der Vleuten, C. P. M., & Berger, M. P. F. (2011). Exploring task- and student-related factors in the method of propositional manipulation (MPM). *Journal of Statistics Education, 19,* 1–23.

Leppink, J., Broers, N. J., Imbos, T., 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.

Leppink, J., Broers, N. J., Imbos, T., 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.

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., Ayres, P., & Pachman, M. (2008). Assessment of cognitive load in multimedia learning environments: Theory, methods, and applications. In D. H. Robinson & G. J. Schraw (Eds.), *Recent innovations in educational technology that facilitate student learning* (pp. 11–35). Charlotte: Information Age.

Paas, F., Renkl, A., & Sweller, J. (2003a). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38,* 1–4.

Paas, F., Tuovinen, J., Tabbers, H., & Van Gerven, P. (2003b). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38,* 63–71.

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

Paas, F., Van Merriënboer, J., & Adam, J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills, 79,* 419–430.

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.

Salden, R. J. C. M., Paas, F., Broers, N. J., & Van Merriënboer, J. J. G. (2004). Mental effort and performance as determinants for the dynamic selection of learning tasks in Air Traffic Control training. *Instructional Science, 32,* 153–172.

Salomon, G. (1984). Television is "easy" and print is "tough": The differential investment of mental effort in learning as a function of perceptions and attributes. *Journal of Educational Psychology, 78,* 647–658.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous and germane cognitive load. *Educational Psychology Review, 22,* 123–138.

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load as a factor in the structuring of technical material. *Journal of Experimental Psychology. General, 119,* 176–192.

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

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

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.

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

Van Gog, T., & Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction, 20,* 95–99.

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.

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.

Zumbach, J., & Mohraz, M. (2008). Cognitive load in hypermedia reading comprehension: Influence of text type and linearity. *Computers in Human Behavior, 24,* 875–887.