cogent education

EDUCATIONAL ASSESSMENT & EVALUATION | RESEARCH ARTICLE

Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence

Moritz Krell[1]*

Abstract: This study evaluates a 12-item instrument for subjective measurement of mental load (ML) and mental effort (ME) by analysing different sources of validity evidence. The findings of an expert judgement (N = 8) provide *evidence based on test content* that the formulation of the items corresponds to the meaning of ML and ME. An empirical study was conducted in which secondary school students (N = 602) worked on multiple choice (mc)-tasks and thereafter using the developed instrument to self-report ML and ME. The findings show that the instrument reliably measures the two positively correlated constructs ML and ME (*evidence based on internal structure*). Students working on mc-tasks with high complexity self-reported higher amounts of ML and ME than students working on mc-tasks with low complexity, and there is a negative relation between test performance and ML (*evidence based in relation to other variables*). Implications for educational assessment and limitations of the study are discussed.

Subjects: Psychometrics/Testing & Measurement Theory; Test Development, Validity & Scaling Methods; Educational Research

Keywords: cognitive load; mental load; mental effort; validation

Received: 28 November 2016
Accepted: 05 January 2017

*Corresponding author: Moritz Krell, Biology Education, Freie Universität Berlin, Schwendenerstraße 1, 14195 Berlin, Germany
E-mail: moritz.krell@fu-berlin.de

Reviewing editor:
Sammy King Fai Hui, The Education University of Hong Kong, Hong Kong

Additional information is available at the end of the article

ABOUT THE AUTHOR

Moritz Krell is a lecturer and a researcher in biology education. His research areas relate to students' and teachers' competencies in scientific inquiry and scientific reasoning with a special focus on modelling competencies. One research area relates to the development and evaluation of assessment instruments including the analysis of difficulty-generating task characteristics. The questionnaire, which is introduced in this study, was developed within this line of research because it allows to receive information about the cognitive demand of assessment instruments.

PUBLIC INTEREST STATEMENT

Cognitive load refers to an individual's cognitive capacity which is used to work on a task, to learn or to solve a problem. Therefore, the measurement of cognitive load can provide an insight into the cognitive demand of tasks. This is useful in educational and psychological settings for several reasons. For instance, high cognitive load may hinder understanding and, therefore, should be considered when developing instructional designs or assessment tasks. This study introduces and evaluates a questionnaire to measure cognitive load in educational assessments. The findings suggest that the questionnaire allows to measure two theoretically established dimensions of cognitive load (mental load and mental effort). Thus, the questionnaire may be used by researchers and practitioners to evaluate the cognitive demand of tasks and, thereby, better understand learners' test performances..

1. Introduction

Cognitive load (CL) can be broadly defined as a multidimensional construct representing an individual's cognitive capacity which is used to work on a task, to learn or to solve a problem (Chandler & Sweller, 1991; Paas, Tuovinen, Tabbers, & Van Gerven, 2003; Paas & Van Merriënboer, 1994; Sweller, Ayres, & Kalyuga, 2011). CL has a causal dimension which reflects the interaction between person- and task-characteristics as well as an assessment dimension which describes the measurable aspects *mental load* (ML), *mental effort* (ME) and *performance* (PE; Paas & Van Merriënboer, 1994). ML is said to be task-related, indicating the cognitive capacity which is needed to process the complexity of a task. In contrast, ME is subject-related and reflects an individual's invested cognitive capacity while working on a task. Sweller et al. (2011) propose ML and ME being two different but, in most cases, positively correlated constructs. De Jong (2010) critically discusses that PE is sometimes conceptualised as being one aspect of CL (e.g. Paas & Van Merriënboer, 1994) and sometimes as being an indicator for CL (e.g. Kirschner, 2002). Furthermore, the relation between PE and CL is not clear. For example, subjects may reach the same number of correct answers in a test (i.e. PE) but need to working with different amounts of ME (Paas et al., 2003).

Measuring CL has become relevant in educational and psychological research due to several reasons. For instance, high CL may hinder understanding and knowledge construction and therefore should be considered when developing instructional designs (e.g. Kirschner, 2002; Kirschner, Sweller, & Clark, 2006; Sweller, van Merrienboer, & Paas, 1998). Furthermore, CL is measured as a control variable in educational assessment in order to better understand task difficulty and students' test performance (e.g. Krell & Tieben, 2014; Nehring, Nowak, Upmeier zu Belzen, & Tiemann, 2012; Poehnl & Bogner, 2013). In addition, findings of CL measures can contribute to a further development of CL theory (Paas et al., 2003).

The present study focuses on subjective measurement of CL. Therefore, subjects are asked to self-report the amount of CL after working on a task (Sweller et al., 2011). Many researchers use this approach (e.g. Krell & Tieben, 2014; Nehring et al., 2012; Paas, 1992; Poehnl & Bogner, 2013) and Paas et al. (2003) emphasise that subjective measures were shown to be reliable and valid. However, subjective measurement of CL "has become highly problematic" (Kirschner, Ayres, & Chandler, 2011, p. 104) due to several reasons (de Jong, 2010; Krell, 2015; Leppink, Paas, Van der Vleuten, Van Gog, & Van Merriënboer, 2013; Paas et al., 2003; van Gog & Paas, 2008):

(1) Many researchers adapt a scale initially developed by Paas (1992) and change the wording or number of category labels without re-evaluating the instrument.

(2) Often, only one single item is used to measure CL, although the use of several items would increase measurement precision.

(3) Sometimes, it is not entirely clear which trait items are aimed to measure. For example, many researchers use category labels related to task complexity but label them broadly as measures of CL.

(4) Finally, van Gog and Paas (2008) criticise that "all measures […] provide indications of cognitive load as a whole rather than of its constituent aspects" (p. 18).

Kirschner et al. (2011) call the development of instruments which separately measure aspects of CL "the holy grail" of CL research but the authors "seriously doubt whether this is possible" (p. 104). Despite this concern, Leppink et al. (2013) proposed an instrument to separately measure content-related (intrinsic load), instruction-related (extraneous load) and process-related (germane load) sources of CL (cf. Paas & Van Merriënboer, 1994). Paas et al. (2003) underline that "cognitive load can be assessed by measuring mental load, mental effort, and performance" (p. 66).

The present study contributes to the issues sketched out above by evaluating an instrument to measure ML and ME by analysing different sources of validity evidence. Validity is a fundamental requirement for the interpretation of empirical research findings (Kane, 2006, 2013; Linn, 2010). In the *Standards for Educational and Psychological Testing*, it is emphasised that "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 11). The authors further elaborate on different "sources of evidence that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use" (p. 13). These sources of validity evidence are: Evidence based on test content, on response processes, on internal structure and on relations to other variables (American Educational Research Association [AERA] et al., 2014). Since validation depends on the intended interpretation and use of test scores, Kane (2013) argues that making the "interpretation/use argument" transparent is an integral part of validation.

The interpretation/use argument in the present case is as follows. The intended use of the instrument called *Students' Mental Load and Mental Effort in Biology Education-Questionnaire* ("StuMMBE-Q") is to provide measures of students' ML and ME as control variables in biology education research. Hence, students' scores on the StuMMBE-Q are interpreted as indicators for the amount of ML and ME while processing given tasks. A basic prerequisite for this purpose is that the content of the StuMMBE-Q (i.e. wording of the items) appropriately represents the constructs ML and ME (*evidence based on test content*). Furthermore, the instrument should provide distinct measures of the two dimensions ML and ME (*evidence based on internal structure*). Since ML and ME are conceptualised to enhance with increasing task complexity (Paas & Van Merriënboer, 1994; Sweller et al., 2011), the StuMMBE-Q should provide higher measures of ML and ME when subjects work on tasks with high complexity than when subjects work on tasks with low complexity (*evidence based on relations to other variables*). Finally, a positive relation between PE and ML and a negative one between PE and ME may be interpreted as additional *evidence based on relations to other variables*, since this source of evidence may be obtained by analysing criteria the respective testing instrument is expected to predict (AERA et al., 2014). However, the relation between ML, ME and PE is theoretically not clearly established (de Jong, 2010; Kirschner, 2002; Krell, 2015; Paas et al., 2003). The fourth source of validity evidence proposed in the *Standards for Educational and Psychological Testing*, which is *evidence based on response processes*, is not considered as part of the validity argument in this study. This is in line with AERA et al. (2014), since no "explicit claims about response processes are made" (p. 16).

2. Method

2.1. Development of the StuMMBE-Q

Based on CL theory and the instrument used by Nehring et al. (2012), who measured CL as one global construct, the StuMMBE-Q was developed consisting of six items representing ML and six items representing ME. For each item, a seven-point rating scale ranging from *not at all* to *totally* was provided (Krell, 2015). The ML-items ask to indicate the complexity of tasks, whereas the ME-items focus on personal effort (Table 1).

Table 1. The 12 items for subjective measurement of ML and ME

Mental load (ML)	Mental effort (ME)
The tasks were difficult to answer [Die Aufgaben waren schwer zu beantworten] (1)	I have put little effort into answering the tasks [Bei der Bearbeitung der Aufgaben habe ich mich wenig bemüht] (2*)
The contents of the tasks were complicated [Der Inhalt der Aufgaben war kompliziert] (3)	I haven't tried hard answering the tasks correctly [Ich habe mir keine besondere Mühe bei der Beantwortung der Aufgaben gegeben] (5*)
The tasks were challenging [Die Aufgaben waren anspruchsvoll] (4)	I have tried hard to answer the tasks correctly [Ich habe mich bei der Bearbeitung der Aufgaben angestrengt] (7)
The tasks were easy to work on [Die Aufgaben waren einfach zu bearbeiten] (6*)	I have made an intellectual effort when answering the tasks [Bei der Beantwortung der Aufgaben habe ich mich geistig angestrengt] (9)
The contents of the tasks were easy to understand [Der Inhalt der Aufgaben war leicht zu verstehen] (8*)	I haven't particularly focused when answering the tasks [Ich habe mich nicht besonders konzentriert, um die Aufgaben zu lösen] (11*)
The tasks were easy to solve [Die Aufgaben waren leicht zu lösen] (10*)	I have given my best to solve the tasks [Ich habe mir Mühe gegeben, um die Aufgaben zu lösen] (12)

Notes: The original version of the StuMMBE-Q is in German language and linguistic flaws may be caused by the translation. Therefore, the German version of each item is provided in brackets. Items with an asterisk were coded reversely. The numbers indicate the position of the items in the StuMMBE-Q. See Krell (2015) for the instrument.

An initial version of the StuMMBE-Q was administered to secondary school students ($N = 188$) in biology classes directly after working on different biology tests ("normal class tests", i.e. no standardised performance measure). This pilot study was used to optimise single items (e.g. their wording). A second pilot study ($N = 506$) was conducted to evaluate the appropriateness of the seven-point rating scale (Krell, 2015). The findings suggested to reduce the scale to a three-point scale allowing to meaningfully distinguish between subjects who report low, medium and high amounts of ML and ME. In this study ($N = 602$), as in the second pilot study, the seven-point scale was *post hoc* reduced to a three-point scale (e.g. Zhu, Updyke, & Lewandowski, 1997). This was done since relevant indices (e.g. rating scale thresholds, point-biserial correlations) proposed that the seven-point scale was not interpreted consistently across the items (Linacre, 2002; Wu, Adams, Wilson, & Haldane, 2007).

2.2. Validity evidence based on test content

Evidence based on test content may be obtained from expert judgements about the relationship between test items and the theoretical construct (AERA et al., 2014; Sireci & Faulkner-Bond, 2014). Therefore, $N = 8$ researchers working in the field of biology education evaluated the items' *domain representation* (Sireci & Faulkner-Bond, 2014) by assigning each item to either ML or ME. The content validity ratio (CVR; Ayre & Scally, 2014; Lawshe, 1975), initially developed to evaluate *domain relevance* (cf. Sireci & Faulkner-Bond, 2014), was adapted to quantify agreement between the experts' ratings: $\mathrm{CVR} = \left(n_i - \left(\frac{N}{2}\right)\right)/\frac{N}{2}$; with n_i being the number of experts assigning an item as theoretically intended and N being the total number of experts (i.e. $N = 8$). Consequently, in this study, n_i is a dichotomous variable, with the possible values of 1 (item was assigned to either ML or ME as theoretically intended) or 0 (item was not assigned as theoretically intended). This procedure resulted in a mean CVR for the overall test of $\mathrm{CVR}_{\mathrm{mean}} = 0.979$, since the eight experts assigned 11 items as theoretically intended (i.e. $\mathrm{CVR}_{\mathrm{item}} = 1$ for 11 items) and only one expert assigned one item (item (11*); see Table 1) not as theoretically intended ($\mathrm{CVR}_{\mathrm{item}} = 0.750$ for item (11*)).

2.3. Sample and performance measure

To obtain *evidence based on internal structure* and *evidence based on relation to other variables*, an empirical study was conducted. The StuMMBE-Q was administered to a sample of 602 students (school years 9 and 10; aged 13–18; 52% female) after working on a standardised multiple choice (MC) test measuring competencies in biological experimentation which served as PE measure (cf. Krell & Vierarm, 2016; Phan, 2007). In these MC tasks, different experimental contexts (e.g. photosynthesis, seed germination) are described and for each context two parallel tasks, with high and low complexity, respectively, exist. In the tasks with high complexity, subjects have to understand and solve problems related to biological experiments with two independent variables, while experiments with only one independent variable are considered in the tasks with low complexity. Hence, task complexity was systematically and objectively varied during task development. In addition, task complexity was empirically shown to be a significant difficulty generating characteristic of the MC tasks (Krell & Vierarm, 2016). In the present study, the students got different test booklets containing either MC tasks with high complexity only, MC tasks with both high and low complexity, or MC tasks with low complexity only.

2.4. Data analysis

Data analysis was done within the framework of item response theory (Bond & Fox, 2001; Embretson & Reise, 2000) using the software ConQuest 3 (Wu et al., 2007). Specifically, the rating scale model (RSM) was applied to analyse students' self-reported ML and ME since this model was shown to be appropriate in the given context (Krell, 2015). For estimating the students' PE in the MC test, the one parametric logistic test model was applied ("Rasch-Model"; Embretson & Reise, 2000). Weighted likelihood estimates (WLE; Wu et al., 2007) were used as estimates for the students' ML, ME and PE.

To provide *evidence based on internal structure*, a one- (1D) and a two-dimensional (2D) RSM have been specified and compared. In the 1D-RSM, a global latent dimension (CL) is assumed, whereas two latent dimensions (ML, ME) are postulated in the 2D-RSM. For the evaluation of the data's internal structure, the model fits of these two models were compared (Rios & Wells, 2014). On item level, ConQuest provides the weighted and unweighted mean of squared standardised residuals (wMNSQ and uMNSQ), which both have an expected value of 1 with, for polytomous IRT-models, a range from 0.6 to 1.4 indicating an acceptable model fit (Wright & Linacre, 1994). Further, the estimated rating scale thresholds (λ_s) should increase monotonically (Krell, 2012; Linacre, 2002). Person (rel.$_{EAP/PV}$) and item reliability (rel.$_{it}$) measures indicate the separability (i.e. stability) of estimated person and item parameters (Bond & Fox, 2001). The relative model fit was analysed using descriptive information indices (i.e. AIC, BIC) as well as the likelihood difference test (LD-test; Krell, 2012; Wu et al., 2007).

To provide *evidence based on relations to other variables*, the three kinds of test booklets were used to define a group variable "task complexity" (test booklets containing MC tasks with low complexity only =0; test booklets with "medium complexity" containing MC tasks with low and high complexity =1; test booklets containing MC tasks with high complexity only =2). This group variable was used as predictor variable in a latent regression of task complexity on ML and ME (Wu et al., 2007). In this analysis, dummy coding was applied using low complexity (=0) as the baseline group against which medium complexity (=1) and high complexity (=2) were compared. Additionally, Pearson correlations between the students' PE and ML as well as PE and ME were calculated.

3. Results

3.1. Validity evidence based on internal structure

The 1D-RSM results in slightly better fit statistics on item level than the 2D-RSM, but these statistics are good for both models compared. The reliability measures also indicate an acceptable fit of both models (Table 2).

The comparison of the models based on relative fit statistics indicates a significant better fit of the 2D-RSM (Table 3). Using this model for estimating WLE, the students self-reported a significantly smaller amount of ML ($M_{WLE(ML)} = -1.719$, $SD_{WLE(ML)} = 1.677$) than of ME ($M_{WLE(ME)} = 0.335$, $SD_{WLE(ME)} = 1.617$, $p < 0.001$, $d = 1.246$). The latent correlation between both dimensions is positive but rather small ($r_{ML/ME} = 0.168$).

3.2. Validity evidence based on relations to other variables

The WLE, indicating the students' self-reported amount of ML and ME, vary (increase) with the complexity of the MC tasks (Figure 1). Consequently, task complexity turns out to be a significant predictor of ML and ME in the 2D-latent regression model (Table 4). As expected, the effect of high complexity on ML and ME is larger than the effect of medium complexity. However, the effect of task complexity on ML is higher than on ME since there is no significant difference in students' self-reported ML between low complexity and medium complexity.

Table 2. Absolute fit statistics for the RSMs

Model	uMNSQ	wMNSQ	λ_s	rel.$_{it}$	rel.$_{EAP/PV}$
1D	0.78–1.14	0.79–1.21	0	0.99	0.74
2D	0.74–1.30	0.82–1.25	0	0.99	0.79/0.77

Notes: The number of items with unordered values of λ_s is given in the column λ_s. For the 2D-model, rel.$_{EAP/PV}$ is provided for both dimensions (ML/ME).

Table 3. Relative fit statistics for the RSM

Model	Parameters	Deviance	AIC	BIC	LD-test
1D	14	12,637	12,665	12,727	759(2) $p < 0.001$
2D	16	11,877	11,909	11,980	

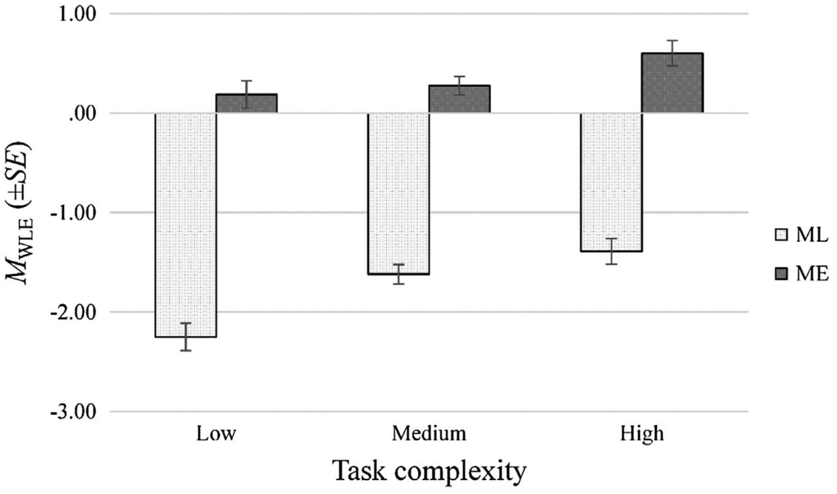

Figure 1. The students' ML and ME ($M_{WLE} \pm SE$), separately shown for task complexity.

Notes: Low = test booklets containing MC tasks with low complexity only; medium = test booklets containing MC tasks with low and high complexity; high = test booklets containing MC tasks with high complexity only.

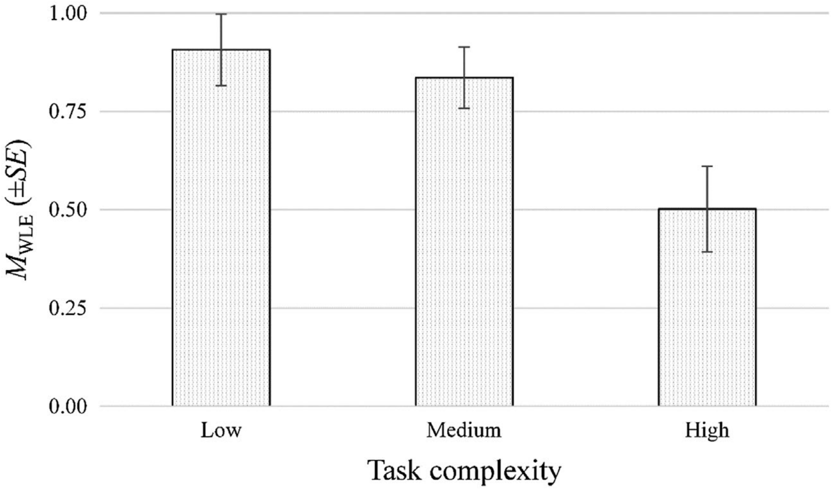

Figure 2. The students' PE ($M_{WLE} \pm SE$), separately shown for task complexity.

Notes: Low = test booklets containing MC tasks with low complexity only; medium = test booklets containing MC tasks with low and high complexity; high = test booklets containing MC tasks with high complexity only.

Table 4. Latent regression of task complexity on ML and ME

		Mental load (ML)			Mental effort (ME)		
		b	***se*(*b*)**	***p***	***b***	***se*(*b*)**	***p***
Constant		−2.443	0.162	<0.001	0.106	0.124	0.196
Task complexity	Medium (=1)	0.441	0.188	0.009	0.133	0.152	0.191
	High (=2)	1.362	0.208	<0.001	0.546	0.172	0.001

Notes: Dummy coding was applied using low complexity (=0) as the baseline group against which medium complexity (=1) and high complexity (=2) was compared.

As ML and ME do, the students' PE in the MC tasks also varies (decreases) with task complexity (Figure 2). Accordingly, there is a significant negative Pearson correlation between PE and ML ($r_{PE/ML} = -0.220$, $p < 0.001$). Thus, the better the students scored in the MC test, the lower was their self-reported amount of ML (and vice versa). No significant correlation between PE and ME was found ($r_{PE/ME} = -0.017$, $p = 0.680$).

4. Conclusion and discussion

As emphasised in the introduction, the development of instruments to separately measure aspects of CL is seen as the "the holy grail" of CL research (Kirschner et al., 2011). Especially, the use of instruments for subjective measurement of CL has become problematical (de Jong, 2010; Kirschner et al., 2011; Krell, 2015; van Gog & Paas, 2008). To contribute to this field of educational and psychological research, this study evaluates the StuMMBE-Q as an instrument to measure ML and ME in biology education. More precisely, it is evaluated to what extent evidence supports the validity of the interpretation of subjects' scores on the StuMMBE-Q as measures of ML and ME as control variables in biology education research. As proposed in the *Standards for Educational and Psychological Testing* (AERA et al., 2014) validity evidence was provided based on test content, on internal structure and on relations to other variables.

As a basic prerequisite for the valid interpretation of the test scores for the intended purpose, the items' content has to appropriately represent the constructs ML and ME (*evidence based on test content*). This evidence may "come from expert judgements of the relationship between parts of the test and the construct" (AERA et al., 2014, p. 14). In this study, researchers working in the field of biology education assigned the items to either ML or ME which resulted in an almost perfect match ($CVR_{mean} = 0.979$; cf. Ayre & Scally, 2014) with the intended formulation of the items. Only one expert assigned item (11*) not as theoretically intended (i.e. $CVR_{item} = 0.750$). However, for $N = 8$, a number of seven agreeing experts (i.e. CVR = 0.750) is proposed to be sufficient for indicating "content validity" (Ayre & Scally, 2014). In addition, the test development provides further *evidence based on test content* since the StuMMBE-Q was developed based on the existing instrument by Nehring et al. (2012) and pilot studies were conducted to evaluate the items' wording and the appropriateness of the rating scale (Sireci & Faulkner-Bond, 2014). However, the present approach for providing "content validity" is rather simple compared to approaches, for example, related to assessment of competencies in science education (e.g. Terzer, Patzke, & Upmeier zu Belzen, 2012). However, this may be justified with the rather low complexity of the present items. For example, the items include no stems and no response alternatives ("distractors") had to be developed and evaluated.

For providing *evidence based on internal structure*, the dimensionality of the data was analysed by evaluating and comparing the fit of two theoretically plausible rating scale models (AERA et al., 2014; Rios & Wells, 2014). The absolute fit statistics propose both models to represent the data well (Table 2) but the relative model comparison supports the 2D-RSM (Table 3). The latent correlation between the dimensions is positive but small. Hence, there is empirical evidence that the StuMMBE-Q allows to reliably measure two positively related but distinct constructs. Taking *evidence based on test content* into account, these constructs are likely to be the students' ML and ME.

In correspondence with the pilot study (Krell, 2015), the seven-point scale was *post hoc* reduced to a three-point scale (Zhu et al., 1997) since data analysis suggested that the seven-point scale was not interpreted consistently across the items. While Leppink et al. (2013) argue that instruments with less than seven response categories would allow measuring on ordinal level only, Stone and Wright (1994, p. 386) emphasise that "more categories do not mean more information" in any case. The present findings suggest that the StuMMBE-Q allows to meaningfully separate students who report low, medium and high amounts of ML und ME.

Paas et al. (2003) emphasise that CL measurement may contribute to further develop CL theory. From this perspective, the present findings suggest not to conceptualise and assess CL as one global construct as it was done, for example, by Nehring et al. (2012), but to separately measure its constituent aspects ML and ME. Thus, in addition to the instrument published by Leppink et al. (2013), the StuMMBE-Q may be used to measure students' ML and ME as control variables in biology education research. Whereas, Leppink et al. (2013) aim to assess content-related (intrinsic load), instruction-related (extraneous load) and process-related (germane load) sources of CL, the present instrument focuses on the perceived complexity of tasks (i.e. ML) and the invested mental effort.

Evidence based on relations to other variables can be provided by "some criteria the test is expected to predict" but also by "group membership variables" (AERA et al., 2014, p. 16). In this study, both kinds of variables were considered. Since ML refers to the cognitive capacity which is needed to process the complexity of a task and ME reflects an individual's invested cognitive capacity when working on a task (Paas & Van Merriënboer, 1994), it is crucial for the intended use of the StuMMBE-Q that students working on tasks with high complexity self-report higher amounts of ML and ME than students working on tasks with low complexity. This evidence could be provided using the ordinal group variable "task complexity" as a predictor variable in a latent regression on ML and ME (Table 4). As additional evidence, Pearson correlations between the students' PE and their self-reported ML and ME were calculated which turned out to be significant but small ($r_{PE/ML}$) and not significant ($r_{PE/ME}$), respectively. These coefficients, therefore, provide only slight validity evidence. However, the non-significant correlation between PE and ME corresponds with findings of other researchers (cf. Kirschner et al., 2011; Sweller et al., 2011) and may be caused, for example, by students working with different amounts of ME but reaching the same score in the MC test (Paas et al., 2003).

Summarising, the findings of this study provide evidence that the formulation of the items corresponds to the theoretical meaning of ML and ME (*evidence based on test content*), that the StuMMBE-Q reliably measures two positively related but distinct constructs which are, thus, likely to be ML and ME (*evidence based on internal structure*), that students working on MC tasks with high complexity self-report higher amounts of ML and ME than students working on MC tasks with low complexity, and that there is a negative relation between students test performance and their reported ML (*evidence based in relation to other variables*).

In addition to the three sources of validity evidence considered in this study, evidence based on response processes is proposed in the *Standards for Educational and Psychological Testing*, but the authors state: "While evidence about response processes may be central in settings where explicit claims about response processes are made [...], there are many other cases where claims about response processes are not part of the validity argument" (AERA et al., 2014, p. 16). However, although not being a central part of the validity argument in this context, evidence for validity based on response processes ("cognitive validity") may be provided in further studies using think-aloud protocols, while respondents answer the StuMMBE-Q (Linn, 2010; Padilla & Benítez, 2014).

As emphasised above, validity cannot be provided per se but only for a particular interpretation of test scores (AERA et al., 2014; Kane, 2006, 2013; Linn, 2010). Therefore, the present instrument may be validly used to provide measures of ML and ME as control variables in biology education research. Further evidence is needed before generalising the findings to include further subjects (e.g. chemistry education; Nehring et al., 2012) or further cognitively challenging situations (e.g. instructional settings; Sweller et al., 1998).

Acknowledgement
The author wants to thank Amelie Vierarm who significantly contributed to data collection.

Funding
The author received no direct funding for this research.

Author details
Moritz Krell[1]
E-mail: moritz.krell@fu-berlin.de
[1] Biology Education, Freie Universität Berlin, Schwendenerstraße 1, 14195 Berlin, Germany.

Citation information
Cite this article as: Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence, Moritz Krell, *Cogent Education* (2017), 4: 1280256.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Ayre, C., & Scally, A. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development, 47*, 79–86. http://dx.doi.org/10.1177/0748175613513808

Bond, T., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

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 de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science, 38*, 105–134. http://dx.doi.org/10.1007/s11251-009-9110-0

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (pp. 17–64). Westport, CT: Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. http://dx.doi.org/10.1111/jedm.2013.50.issue-1

Kirschner, P. (2002). Cognitive load theory: Implications of cognitive load theory on the design of learning. *Learning and Instruction, 12*, 1–10. http://dx.doi.org/10.1016/S0959-4752(01)00014-7

Kirschner, P., Ayres, P., & Chandler, P. (2011). Contemporary cognitive load theory research: The good, the bad and the ugly. *Computers in Human Behavior, 27*, 99–105. http://dx.doi.org/10.1016/j.chb.2010.06.025

Kirschner, P., Sweller, J., & Clark, R. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*, 75–86. http://dx.doi.org/10.1207/s15326985ep4102_1

Krell, M. (2012). Using polytomous IRT models to evaluate theoretical levels of understanding models and modeling in biology education. *Science Education Review Letters, Theoretical Letters, 2012*, 1–5. Retrieved from http://edoc.hu-berlin.de/serl/2012-1/PDF/2012_1.pdf

Krell, M. (2015). Evaluating an instrument to measure mental load and mental effort using item response theory. *Science Education Review Letters, Research Letters, 2015*, 1–6. Retrieved from http://edoc.hu-berlin.de/serl/2015/1/PDF/2015-1.pdf

Krell, M., & Tieben, S. (2014). Goal-Framing in der Kompetenzdiagnostik [Goal-framing in competence assessment]. *Schriftenreihe Fachdidaktische Forschung*, 1–21. Retrieved August 10, 2014, from http://nbn-resolving.de/urn:nbn:de:gbv:hil2-opus4-4050

Krell, M., & Vierarm, A. (2016). Analyse schwierigkeitserzeugender Aufgabenmerkmale bei einem Multiple-Choice-Test zum Experimentieren [Analysis of difficulty generating characteristics of a multiple choice test assessing competencies in experimentation]. In U. Gebhard & M. Hammann (Eds.), *Lehr- und Lernforschung in der Biologiedidaktik* (pp. 283–298). Innsbruck: Studienverlag.

Lawshe, C. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575. http://dx.doi.org/10.1111/peps.1975.28.issue-4

Leppink, J., Paas, F., Van der Vleuten, C., Van Gog, T., & Van Merriënboer, J. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods, 45*, 1058–1072. http://dx.doi.org/10.3758/s13428-013-0334-1

Linacre, J. (2002). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85–106.

Linn, R. (2010). Validity. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (pp. 181–185). Oxford: Elsevier. http://dx.doi.org/10.1016/B978-0-08-044894-7.00893-9

Nehring, A., Nowak, K., Upmeier zu Belzen, A., & Tiemann, R. (2012). Doing inquiry in chemistry and biology: The context's influence on the students' cognitive load. *La Chimica nella Scuola, XXXIV*, 253–258.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill 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. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6*, 351–371. http://dx.doi.org/10.1007/BF02213420

Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*, 136–144.

Phan, T. (2007). *Testing levels of competencies in biological experimentation* (Doctoral dissertation). Christian-Albrechts-Universität Kiel. Retrieved from http://eldiss.uni-kiel.de/macau/receive/dissertation_diss_2130

Poehnl, S., & Bogner, F. (2013). A modified refutation text design: Effects on instructional efficiency for experts and novices. *Educational Research and Evaluation, 19*, 402–425. http://dx.doi.org/10.1080/13803611.2013.789229

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*, 108–116.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*, 100–107.

Stone, M., & Wright, B. (1994). Maximizing rating scale information. *Rasch Measurement Transactions, 8*, 386.

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

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

Terzer, E., Patzke, C., & Upmeier zu Belzen, A. (2012). Validierung von Multiple-Choice Items zur Modellkompetenz durch lautes Denken [Validation of multiple-choice items for the assessment of model competence using thinking aloud protocols]. In U. Harms & F. Bogner (Eds.), *Lehr- und Lernforschung in der Biologiedidaktik* (pp. 45–62). Innsbruck: Studienverlag.

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

Wright, B., & Linacre, J. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

Wu, M. L., Adams, R., Wilson, M., & Haldane, S. (2007). *ACER ConQuest*. Camberwell: ACER Press.

Zhu, W., Updyke, W., & Lewandowski, C. (1997). Post-hoc Rasch analysis of optimal categorization of an ordered-response scale. *Journal of Outcome Measurement, 1*, 286–304.