Science Education
Review Letters

Evaluating an instrument to measure mental load and mental effort using Item Response Theory

Moritz Krell[a*]
[a] Department of Biology, Biology Education, Freie Universität Berlin, Germany

[*] Corresponding author: moritz.krell@fu-berlin.de

Received 24th January 2014, Accepted 12th March 2015

Abstract
Measurement of cognitive load (CL) is seen as a problematic issue since no consensus about appropriate instruments has been reached. In this study, a rating scale instrument to measure mental load (ML; 6 items) and mental effort (ME; 6 items) is evaluated using Item Response Theory. *N*=506 students self-reported their amount of ML and ME after working on a standardised multiple choice-test. The findings propose to separately measure ML and ME instead of CL in general. Furthermore, the 7-point rating scale had to be reduced post-hoc to a 3-point scale in order to reach consistent information. Finally, there was a significant (negative) correlation between ML and test performance, but not between ME and test performance.

Keywords
cognitive load; assessment; mental effort; mental load; performance; rating scale-model

Introduction

Cognitive load (CL) can be broadly defined as an individual's cognitive capacity which is used to work on a task, to learn, or to solve a problem (Sweller et al. 2011). CL theory has become relevant in educational research. For instance, CL may significantly influence a learner's performance and therefore should be considered when developing instructional designs (Sweller et al. 1998).
The present study focuses on subjective measurement of CL. In such an approach, respondents are asked to self-report the amount of CL after working on a task (Sweller et al. 2011). Paas et al. (2003) emphasise that subjective measures were shown to be reliable and valid. Consequently, many researchers use this approach (e.g., Nehring et al. 2012; Paas 1992). However, measurement of CL 'has become highly problematic' (Kirschner et al. 2011, p. 104). With regard to subjective measurement, there are several reasons for this:
(1) Many studies adapt a scale initially developed by Paas (1992) and change the wording or number of category labels without re-evaluating its psychometric properties (Paas et al. 2003; van Gog & Paas 2008).
(2) Often, only a single item is used to measure CL, although the use of several items would increase measurement precision (Leppink et al. 2013).
(3) Sometimes, it is not entirely clear which trait items are aimed to measure. Whereas Paas (1992) focused on mental effort, many research-

ers use category labels related to task complexity but label them broadly as measures of CL (de Jong 2010; van Gog & Paas 2008).
(4) Finally, van Gog and Paas (2008) emphasise 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 to 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) recently published an instrument to separately measure the aspects intrinsic, extraneous, and germane load (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 these issues by evaluating an instrument to measure mental load and mental effort using Item Response Theory (IRT; Bond & Fox 2001) and, by doing so, extends empirical findings in CL research which are in most cases based on Classical Test Theory (CTT).

Theoretical background

Cognitive load
'Cognitive load can be defined as a multidimensional construct representing the load that performing a particular task imposes on the learner's cognitive system' (Paas et al. 2003, p. 64). 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 when 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 the aspects ML and ME 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 different aspects of CL is seen as challenging and the 'the holy grail' of CL research (Kirschner et al. 2011). Furthermore, most instruments are not evaluated due to their category labels used and no consensus has been reached about how many categories are appropriate for meaningfully measuring CL (van Gog & Paas 2008).

Item Response Theory
Compared to CTT, IRT has some important advantages (Bond & Fox 2001). First of all, item parameters (i.e., difficulties) are computed independently from the current sample. Therefore, the calibration of an instrument and the measurement of an attribute are not confounded (Bond & Fox 2001). For rating scales, item-specific parameters are estimated in order to allow for item-specific meanings of category labels (Wei et al. 2014). Furthermore, the appropriateness of the respective IRT-model used can be evaluated based on several fit-indices. More specifically, the adequacy of the number of categories (e.g., Lee & Paek 2014; Zhu et al. 1997) as well as the dimensionality of a given dataset can be evaluated (e.g., Wei et al. 2014). Therefore, IRT is commonly used to evaluate measurement instruments (e.g., Krell 2012; Lee & Paek 2014; Wei et al. 2014).

Aim of this study

This study evaluates an instrument to measure ML and ME using IRT. The following research questions are discussed:
RQ1: To what extent does the instrument distinguishably measure the aspects ML and ME?
H1: Referring to Paas and van Merriënboer (1994) it is assumed to find a two-dimensional structure in the data according to the aspects ML and ME.
RQ2: To what extent is a 7-point rating scale appropriate to measure ML and ME?
H2: As it is done by many researchers (cf., van Gog & Paas 2008), it is expected that a rating scale with seven categories is appropriate.
RQ3: To what extent can evidence be found for the assumption of significant relationships between PE and the aspects ML and ME?
H3: A negative correlation is expected to be found between an individual's PE and the expressed amount of ML, and a positive correlation between PE and ME.

Methodology

Testing instrument
Based on CL theory and the instrument used by Nehring et al. (2012) who measured CL as one global construct, a testing instrument was developed consisting of 6 items representing ML and 6 items representing ME. For each item, a 7-point rating scale was provided. The ML-items ask to indicate the complexity of tasks, whereas the ME-items focus on personal effort (see appendix).
An initial version of the questionnaire 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).

Sample and data collection
The final instrument was administered to a sample of 506 students (school years 9 and 10; aged 13 to 18; 53% female) after working on a standardised multiple choice-test measuring scientific inquiry competencies (cf. Phan 2007) which served as performance measure.

Data analysis
Data analysis was done within the framework of IRT using the software ConQuest 3 (Wu et al. 2007). Specifically, the rating scale-model (RSM; Andrich 1978) was applied. The mean score of five plausible values (m_{PVs}) was used as a measure of each student's response behaviour (Wu et al. 2007).
To discuss *RQ1* and *RQ2*, the model-fit of one- (1D) and two-dimensional (2D) RSMs have been compared. In the 1D-RSMs, an overall latent dimension (CL) is assumed, whereas two latent dimensions (ML, ME) are postulated in the 2D-RSMs. On item level, ConQuest provides the weighted and unweighted mean of squared standardised residuals (wMNSQ and uMNSQ; Wu et al. 2007) 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 (Bond & Fox 2001). Further, item-specific thresholds (δ_{is}) and m_{PVs} should increase monotonically across the response categories to support the assumption of an (at least) ordinal scale (Krell 2012; Wu et al. 2007). Person- ($rel._{EAP/PV}$) and item-reliability ($rel._{it}$) measures indicate the separability (i.e. stability) of person and item parameters estimated in the respective RSM (Bond & Fox 2001). The relative model-fit was analysed using descriptive information indices (i.e. AIC, BIC) and the chi square difference test (χ^2-test; Wu et al. 2007). To discuss *RQ3*, Pearson correlations between m_{PVs} and PE were analysed.

Results

Both the 1D- and the 2D-RSMs resulted in poor item-fit parameters (Tab. 1). For example, the thresholds δ_{is} did not increase monotonically in 12 of 12 items. Therefore, to optimise the estimation, the response categories were reduced post-hoc (cf. Zhu et al. 1997). First, the response categories 1/2 and 6/7 were combined to reach a 5-point scale. As this still resulted in items with poor fit parameters (Tab. 1), a 3-point scale was created post-hoc by combining the categories 1/2, 3/4/5, and 6/7.

Tab. 1. Absolute fit statistics for the different RSMs

scale	model	uMNSQ	wMNSQ	δ_{is}	m_{PVs}	$rel._{it}$	$rel._{EAP/PV}$
7-point	1D	0.7 to 1.3	0.7 to 1.3	12	10	.99	.76
	2D	0.7 to 1.4	0.7 to 1.3	12	6	.97	.79 / .82
5-point	1D	0.8 to 1.2	0.8 to 1.3	12	5	.99	.75
	2D	0.7 to 1.4	0.8 to 1.3	12	0	.97	.76 / .78
3-point	1D	0.8 to 1.2	0.8 to 1.2	0	1	.99	.73
	2D	0.7 to 1.3	0.8 to 1.2	0	0	.96	.77 / .75

Note. The number of items with unordered values of δ_{is} and m_{PVs} are given in the columns δ_{is} and m_{PVs}. For the 2D-models, $rel._{EAP/PV}$ is provided for both dimensions (ME / ML).

The 3-point 2D-RSM shows reasonable good item-fit statistics: The MNSQs range from 0.7 to 1.3 and the values of δ_{is} and m_{PVs} increase monotonically across the categories in all items. The item separation reliability is good. The variance of the students' responses in both dimensions is $var._{ML}$=2.94 and $var._{ME}$=1.84 which is sufficiently large to allow for an acceptable

separation of the persons' PVs in both dimensions (Tab. 1).
The information indices also suggest the 3-point scale models as best fitting. Based on the χ^2-test, the 2D-RSM fits significantly better than the 1D-RSM for each scale length (Tab. 2).

Tab. 2. Relative fit statistics for the different RSMs

scale	model	parameters	deviance	AIC	BIC	χ^2-test
7-point	1D	18	20,565	20,601	20,677	796.07(2);
	2D	20	19,769	19,809	19,893	$p < .000$
5-point	1D	16	16,170	16,202	16,270	676.07(2);
	2D	18	15,494	15,530	15,606	$p < .000$
3-point	1D	14	10,645	10,673	10,732	622.43(2);
	2D	16	10,022	10,054	10,122	$p < .000$

In the 2D-RSM, the latent correlation between ML and ME is positive but small (r=.18). On average, the students reported a significantly smaller amount of ML (m_{PVs}=-1.84, sd=1.54) than of ME (m_{PVs}=0.43, sd=1.24, d=1.62; large effect). Finally, there is no significant ($r_{PE/ME}$=-.07, p=.11) or significant and medium ($r_{PE/ML}$=-.39, p<.01) Pearson correlation between the students' PVs and PE.

Discussion and conclusion

It was assumed that a two-dimensional structure would be found in the data, according to ML and ME. This assumption can be confirmed, as the 2D-RSMs show a significantly better model-fit than the 1D-RSMs (Tab. 2) and the latent correlation between the dimensions is small. Hence, there is evidence based on internal structure that the present instrument allows inferences to be made about students' ML and ME. These findings propose to separately measure ML and ME instead of measuring CL in general (Sweller et al. 2011; van Gog & Paas 2008).
As also carried out by other researchers, a 7-point scale was provided in the questionnaire (cf. van Gog & Paas 2008). As this scale could not be modelled adequately using the RSM, the 7-point scale was reduced to a 3-point one (cf. Zhu et al. 1997). Leppink et al. (2013) criticises the fact that instruments with less than 7 response categories would not allow measuring on interval but only on ordinal level. However, in this study, 7- and 5-point scales did not provide consistent information (Tab. 1) which could be caused by respondents who are not able to clearly distinguish between adjacent category labels (Lee & Paek 2014). The present findings suggest that the instrument allows separating between students who report low, medium, and high amounts of ML und ME.
A negative correlation between ML and PE and a positive one between ME and PE was expected. Essentially, this hypothesis can be confirmed only partly. For ME, the correlation is not significant, but for ML it is. Therefore, the findings provide evidence to support the theoretical assumptions which conceptualise a causal relationship between ML and PE, but not for a relationship between ME and PE (Paas & van Merriënboer 1994). This corresponds with findings of other researchers (cf. Kirschner et al. 2011) and may be caused, for example, by individuals who reach the same number of correct answers in a test but are working with different amounts of ME (Paas et al. 2003). Therefore, 'estimates of mental effort may yield important information that is not necessarily reflected in mental load and performance measures' (Paas et al. 2003, p. 65).
Kirschner et al. (2011) called the development of instruments to separately measure aspects of CL the 'holy grail' of CL research. In addition to the instrument recently put forward by Leppink et al. (2013), the present instrument may be used to measure students' ML and ME. 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 and the invested mental effort. In the present study the students reported a significantly smaller amount of ML than of ME. Such information may be used to further investigate relations between perceived task difficulty and students' motivation to investigate ME to complete a given task (cf. van Gog & Paas 2008).

References

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*, 581–594.

Bond, T. & Fox, C. (2001). *Applying the Rasch model*. Mahwah, NJ: Erlbaum.

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design. *Instructional Science, 38*, 105–134.

Kirschner, P., Ayres, P. & Chandler, P. (2011). Contemporary cognitive load theory research. *Computers in Human Behavior, 27*, 99–105.

Kirschner, P. (2002). Cognitive load theory. *Learning and Instruction, 12*, 1–10.

Krell (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.

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*. doi:10.3758/s13428-013-0334-1

Nehring, A., Nowak, K., Upmeier zu Belzen, A. & Tiemann, R. (2012). Doing inquiry in chemistry and biology. *La Chimica nella Scuola, XXXIV*, 253–258.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics. *Journal of Educational Psychology, 84*, 429–434.

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.

Paas, F. & van Merriënboer, J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6*, 351–371.

Phan, T. (2007). *Testing levels of competencies in biological experimentation* (Doctoral dissertation). Christian-Albrechts-Universität, Kiel.

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

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

van Gog, T. & Paas, F. (2008). Instructional efficiency. *Educational Psychologist, 43*, 16–26.

Wei, S., Liu, X. & Jia, Y. (2014). Using Rasch measurement to validate the instrument of Students' Understanding of Models in Science (SUMS). *International Journal of Science and Mathematics Education, 12*, 1067–1082.

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

Appendix

Note that the original version of the instrument is in German language and that linguistic flaws may be caused by the translation. Therefore, for the sake of transparency, the German version of each item is provided in brackets. Items with an asterisk were coded reversely (i.e. 1→7, 2→6, ..., 7→1).

Your opinion is in demand!

For the end I would like to learn how difficult you found the just finished test altogether. So refer at the reply to the following questions to the test in the whole, not to single tasks.
Of course, your answers to the following questions aren't graded.

Please indicate with one *X* on each line how strongly the following statements are true for you.	not at all (1)	(2)	(3)	moderately (4)	(5)	(6)	totally (7)
The tasks were difficult to answer. [Die Aufgaben waren schwer zu beantworten.]	☐	☐	☐	☐	☐	☐	☐
At the processing of the tasks I haven't done my best particularly. * [Bei der Bearbeitung der Aufgaben habe ich mich wenig bemüht.]	☐	☐	☐	☐	☐	☐	☐
The contents of the tasks were complicated. [Der Inhalt der Aufgaben war kompliziert.]	☐	☐	☐	☐	☐	☐	☐
The tasks were challenging. [Die Aufgaben waren anspruchsvoll.]	☐	☐	☐	☐	☐	☐	☐
I haven't taken particular trouble with the reply to the tasks. * [Ich habe mir keine besondere Mühe bei der Beantwortung der Aufgaben gegeben.]	☐	☐	☐	☐	☐	☐	☐
The tasks were easy to work on. * [Die Aufgaben waren einfach zu bearbeiten.]	☐	☐	☐	☐	☐	☐	☐
I have made an effort at the processing of the tasks. [Ich habe mich bei der Bearbeitung der Aufgaben angestrengt.]	☐	☐	☐	☐	☐	☐	☐
The contents of the tasks were easy to understand. * [Der Inhalt der Aufgaben war leicht zu verstehen.]	☐	☐	☐	☐	☐	☐	☐
At the reply to the tasks I have made an effort intellectually. [Bei der Beantwortung der Aufgaben habe ich mich geistig angestrengt.]	☐	☐	☐	☐	☐	☐	☐
The tasks were easy to solve. * [Die Aufgaben waren leicht zu lösen.]	☐	☐	☐	☐	☐	☐	☐
I haven't particularly focused me to solve the tasks. * [Ich habe mich nicht besonders konzentriert, um die Aufgaben zu lösen.]	☐	☐	☐	☐	☐	☐	☐
I have given my best to complete the tasks. [Ich habe mir Mühe gegeben, um die Aufgaben zu lösen.]	☐	☐	☐	☐	☐	☐	☐

Thank you!