DEVELOPMENT AND VALIDATION OF THE MCLS-POL The Validation and Further Development of the Multidimensional Cognitive Load Scale for Physical and Online Lectures (MCLS-POL) DEVELOPME...

Article *in* Frontiers in Psychology · March 2021

DOI: 10.3389/fpsyg.2021.642084

CITATIONS
0

READS
18

2 authors:

Martin Stolpe Andersen
University of Copenhagen
3 PUBLICATIONS **1** CITATION
SEE PROFILE

Guido Makransky
University of Copenhagen
68 PUBLICATIONS **1,471** CITATIONS
SEE PROFILE

The Validation and Further Development of the Multidimensional Cognitive Load Scale for Physical and Online Lectures (MCLS-POL)

Martin S. Andersen[1] and Guido Makransky[1]

[1] Department of Psychology, University of Copenhagen, Copenhagen, Denmark

Corresponding Author: Martin S. Andersen, University of Copenhagen

Address: Øster Farimagsgade 2A, 1353 København K

Abstract

Cognitive load theory (CLT) has been widely used to help understand the process of learning and to design teaching interventions. The Cognitive Load Scale (CLS) developed by Leppink et al., (2013) has emerged as one of the most validated and widely used self-report measures of intrinsic load (IL), extraneous load (EL), and germane load (GL). In this paper we investigated an expansion of the CLS by using a multidimensional conceptualization of the EL construct that is relevant for physical and online teaching environments. The Multidimensional Cognitive Load Scale for Physical and Online Lectures (MCLS-POL) goes beyond the CLS's operationalization of EL by expanding the EL component which originally included factors related to instructions/explanations with sub-dimensions including EL stemming from noises, and EL stemming from both media and devices within the environment. Through three studies, we investigated the reliability, and internal and external validity of the MCLS-POL using the Partial Credit Model, Confirmatory Factor Analysis, and differences between students either attending a lecture physically or online (Study 2 and 3). The results of Study 1 (N = 250) provide initial evidence for the validity and reliability of the MCLS-POL within a higher education sample, but also highlighted several potential improvements which could be made to the measure. These changes were made before re-evaluating the validity and reliability of the measure in a new sample of higher education psychology students (N = 140, Study 2), and

psychological testing students (N = 119, Study 3). Together the studies provide evidence for a multidimensional conceptualization cognitive load and provide evidence of the validity, reliability, and sensitivity of the MCLS-POL and provide suggestions for future research directions.

Keywords: cognitive load, confirmatory factor analysis, item response theory, online lectures, Rasch measurement

Introduction

Cognitive load theory (CLT) posits that the strain put on working memory by the learning content plays a key role in whether or not the student succeeds in learning (Sweller et al., 2011a). A fundamental assumption is that working memory is limited in terms of capacity, but long term memory has a much greater capacity as information is stored in schemas (Chi et al., 1982). Thus, working memory becomes a form of bottleneck that requires instructors to design learning content in a way that can maximize the amount of information that is stored in the long term memory.

Originally, the CLT assumed that cognitive load was a unidimensional construct pertaining only to the total capacity of working memory (Ayres, 2018), and there is still disagreement about how to conceptualize different types of cognitive load (Kalyuga, 2011 Tindall-Ford et al., 2019). However, three distinct types of cognitive load are primarily described, including: Intrinsic Load (IL), which relates to the students perceived difficulty of the learning material, the difficulty of the learning material varies based on the materials composition and the materials' element interactivity (Sweller et al., 1998; Tindall-Ford et al., 2019). Extraneous Load (EL) which consist of non-intrinsic parts of the learning situation i.e.

non relevant information presented together with relevant information or inefficient instructional design, which will unnecessarily strain the working memory of the student (Sweller et al., 2011b). Finally, Germane Load (GL) is already existing cognitive resource which can ease the learning e.g. strategies for learning (Ayres, 2018; Sweller et al., 2011b). Some researchers have argued that GL is part of IL (Kalyuga, 2011; Sweller, 2010). Others argue that it makes sense to separate GL from IL and describe how GL is tied to actual effort that leads to a better understanding of the content (e.g., Klepsch et al., 2017), Finally, a recent article by Klepsch and Seufert argues that IL stems from a passive experience of a task, opposite GL that stems from an active experience of a task. (Klepsch & Seufert, 2021).

Many attempts at measuring cognitive load have been proposed including objective tasks such as secondary tasks (Sweller et al., 2011c) and psychophysiological measures such as eye tracking (Scharinger et al., 2020; Zheng & Cook, 2012), and EEG (Antonenko et al., 2010; Baceviciute et al., 2020; Makransky, Terkildsen, et al., 2019). Recently an article by Minkley, Xu and Krell have compared subjective and objective factors of CL which found heart rate to be related to self-reported metal effort but not self-reported mental load, and self-reported mental effort and mental load predicted task performance better than heart rate measures (Minkley et al., 2021) . However, the most common way to measure cognitive load is through self-report measures.

Previously, a single item measure by Paas (1992) has been widely used and further developed to measure several types of cognitive load (Ayres, 2006; Cierniak et al., 2009). However, single item scales have also been criticized due to several limitations including being too simplistic, making it difficult for learners to make sensible distinctions between the complexity of the material (IL) and inadequate instructions (EL; Kirschner et al., 2011). Several

other self-report scales assess cognitive load with multiple items including a scale developed by Klepsch and colleagues (2017) which assess IL, EL and GL, a measure to assess mental load and mental effort developed by Krell and colleagues (2017), in addition to the cognitive load scale by Leppink and colleagues(2013), which we use in this article. We have chosen to build on the cognitive load scale by Leppink and colleagues as the items assess a broader domain such as a lecture.

In this article we aim to validate a revised version of Leppink and colleagues' Cognitive Load Scale (Leppink et al, 2013; CLS). The CLS has been widely used in educational settings, and several studies provide support for the validity and reliability of the instrument (e.g., Andersen & Makransky, 2021; Hadie & Yusoff, 2016; Leppink et al., 2013). This includes construct validity assessed through exploratory factor analysis (Leppink et al., 2013) or confirmatory factor analyses (Andersen & Makransky, 2021; Hadie & Yusoff, 2016; Leppink et al., 2013) and item response theory (Andersen & Makransky, 2021). The reliability has also been examined, typically through Cronbach's Alpha (Cronbach, 1951) or similar estimates (Andersen & Makransky, 2021; Hadie & Yusoff, 2016; Leppink et al., 2013) and furthermore the external validity has been examined by investigating how the scales are correlated to learning outcomes (Andersen & Makransky, 2021). Although there is mounting evidence of the reliability and construct validity of the CLS, there are still several gaps in this literature. A main gap in the literature are that there may be a need to revisit the content validity of the EL dimension of the CLS, and there is a need to evaluate the sensitivity of these potential dimensions of EL in physical and online lectures.

Regarding the content validity of the EL dimension, a recent study suggests that the EL may be a multidimensional construct consisting of several sub-components. Andersen and

Makransky (2021) provide reliability and validity evidence that the EL dimension should be split into three subscales measuring distinct forms of EL in virtual reality environments. The subscales included: EL stemming from instructions (e.g. "The instructions and/or explanations used in the simulation were very unclear."), EL stemming from interaction (e.g. "The interaction technique used in the simulation made it harder to learn."), and EL stemming from the environment (e.g. "The virtual environment was full of irrelevant content."). This multidimensional conceptualization was theorized within immersive environments, but has not been suggested or investigated in traditional teaching environments. In this article we propose that the multidimensional conceptualization of EL is not only relevant in virtual learning environments, but rather that it is also necessary for accurately measuring cognitive load in physical and online lectures. Although, cognitive load theory does not clearly address the idea that disturbances and noises might increase EL (Sweller et al., 2011b), research suggests that multitasking using mobile devices reduce learning (Chen & Yan, 2016; Kuznekoff & Titsworth, 2013). Furthermore, research suggests that noises in learning environments can also influence learning (Ali, 2013; Servilha et al., 2014), thus the idea that noises and disturbances add to EL seems straightforward. Therefore, we have devised items for three subscales to measure EL in relation to physical and online lectures addressing contemporary issues, including noises in the environment or distractions from devices such as mobile phones, which might provoke EL. In addition to the original conceptualization of EL from Leppink et al., (2013) that includes instructions and or explanations (e.g. "*The instructions and/or explanations during the activity were very unclear*"), our theoretical conceptualization of EL includes sub-dimensions stemming from noise (e.g. "*Noises in the environment made it difficult to focus on the learning content*"), and devices (e.g. "*My activities on my phone/computer made it difficult to focus on the learning*

content"). Besides the newly developed EL subscales we also employed the Intrinsic Load subscale (e.g. "*The topics covered in the activity were very complex*"), and the Germane Load (GL) subscales (e.g. "*The activity really enhanced my understanding of the topic(s) covered*"). The new measure is labeled the Multidimensional Cognitive Load Scale for Physical and Online Lectures (MCLS-POL and can be seen in Table 1). Although several factors influence teaching in both online (Elkaseh et al., 2015) and offline learning (Kappe & van der Flier, 2012; McKenzie & Schweitzer, 2001) environments, we propose that these components of cognitive load are specifically relevant factors that can influence learning in offline (Cerdan et al., 2018; Chen & Yan, 2016; Klatte et al., 2013) and online lectures (Blasiman et al., 2018; Costley et al., 2020; Zureick et al., 2018). Specifically, with the global COVID-19 pandemic, the use of online teaching platforms is quickly increasing (König et al., 2020) and there is evidence that factors such as noise (Servilha et al., 2014) and disturbances from devices (Chen & Yan, 2016) can create cognitive load when learning.

A related gap in the literature that we attempt to account for in this article is the limited number of studies that investigate the sensitivity of the different dimensions of cognitive load in realistic learning environments. Currently some studies have found meaningful differences of groups in cognitive load (Andersen & Makransky, 2021; Klepsch et al., 2017) and others have found predictive validity through regression analyses (Andersen & Makransky, 2021; Zukić et al., 2016).

In this paper we conduct three studies. In the first study, we validate each sub-scale of the MCLS-POL using the Partial Credit Model (PCM) from Item Response Theory (IRT), and second, we used confirmatory factor Analysis (CFA) to investigate the structural validity of the MCLS-POL. In the second and third studies we implement changes to the scale and investigate

the sensitivity of the different sub-dimensions during a lecture in a higher education psychology bachelor course on the topic of educational psychology (Study 2), and a different lecture in a higher education psychology masters course on the topic of psychological testing (Study 3) which both took place in the Fall 2020 semester. Importantly, the setting of Study 2 and Study 3 took place during the COVID-19 pandemic and students were selected to either attend the lecture in person or online via Zoom (Kohnke & Moorhouse, 2020) which gave us the opportunity to investigate if the components of cognitive load differed across settings. Finally, we compared scores across Study 2 and Study 3 to investigate whether the scales would reflect the difference between the two courses. Thus in this article our aim is to investigate whether it is possible to develop and validate questionnaires measuring cognitive load, particularly the expanded scales of extraneous cognitive load pertaining to EL from instruction, noise, and devices. In regard to comparing online with off-line learning, our research hypotheses are that there should be no differences in terms of intrinsic cognitive load or germane cognitive load between online and offline lectures as the materials and the possible germane resources should be similar. Furthermore, we don't expect any differences in relation to EL from instructions, since students in both online and off-line learning environments receive the same instructions. However, we expected to find differences across the newly developed extraneous cognitive load scales related to devices and noises because there will be differences between the online and off-line learning contexts which could influence these factors of EL.

Study 1

Methods Study 1

Sample

Data was collected at a European university during the fall 2019 semester. The psychology students (N = 250) were asked to voluntarily answer a short online survey in relation to their current course in educational psychology (n = 120) or psychological testing (n = 130). A total of 80.8 % reported being females (n = 202), 18.4 % males (n = 46), and 0.8 % (n = 2) reported another gender than male or female. The mean age was 25.46 with an *SD* of 5.45.

Item Development

A team of subject matter experts consisting of an expert in educational psychology, a specialist in human computer interaction, and a psychometrician further developed the scales of Leppink and colleagues' CLS. Based on a previous study where the EL scale was conceptualized using three separate EL subscales aimed at measuring CL in virtual reality (Andersen & Makransky, 2021), we took a similar approach by conceptualizing EL as a multidimensional construct with several subscales. However, instead of being aimed at learning in virtual reality it was aimed at learning during lectures, and was based on the literature that specified the factors that could create extraneous cognitive load within physical and online lectures. We aimed to make the items so generic that it should be possible to transfer them from one context to another without rewriting them, however in keeping with Leppink and colleagues' (2013) formulation item 2 of the Germane Load scale specifically mentions the course subject (i.e. "The activity really enhanced my knowledge and understanding of [course subject].") and therefore has to be modified accordingly for each study (see Table 1 for all items used in Study 1).

Statistical Analyses

In this study, we employ two methodologies to investigate the construct validity of the MCLS-POL. The first methodology is that of item response theory (IRT; Embretson & Reise, 2000) which estimates a probability function for endorsing each item of a scale in relation to the

scales' total score, that allows for detailed analyses of each item. The second methodology is that of confirmatory factor analysis (CFA; Kline, 2011), in which we model the relationship between items and several latent variables called factors. Therefore, we can evaluate the fit of a model including all items and all scales as latent factors and their relation in just one model.

As the IRT approach is focused on each individual scale, it makes sense to conduct the IRT analyses first and let the knowledge from the IRT analyses inform the overall CFA model which will contain all scales. For an IRT model of the Rasch model family to be valid it must live up to five assumption (Rosenbaum, 1989). The assumptions are: (a) unidimensionality; the scale must measure one latent construct only, (b) the items must be monotonic in relation to the total scale, (c) the items must be locally independent, i.e. the items are conditionally independent after accounting for the total score, (d) the items must not show differential item functioning, e.g. students of the same ability should have equal probability of endorsing an item regardless of gender or age, (e) items must be homogenous such that that the rank order of the items of the difficulties remain the same despite differing abilities of the respondent, e.g. the most difficult item should be the most difficult item to endorse for all respondents. We will address each assumption for every scale in the analyses.

In some cases where we find deviations from assumptions of no Differential Item Functioning (DIF) (d) or no Local Dependence (LD) (c), we are still able to obtain close to optimal measurement. When DIF or LD is uniform, we can model this with a graphical log linear Rasch model (GLLRM; Kreiner & Christensen, 2004). This model can account for the differences in item functioning when DIF is present, however when using sum scores we will need to equate across DIF affected groups to make the sums scores comparable. When uniform LD is present it does not influence the sums scores, however LD dependency will inflate

estimates of reliability such as Cronbach's Alpha (Cronbach, 1951) and we will instead use a Monte Carlo method to compute the estimate of reliability (Hamon & Mesbah, 2002). Factor analysis can be used to create a model where each item's relation to the scales is part of a matrix of regressions. In the confirmatory approach, we restrict the model, such that items of a given scale only load on the hypothesized factor and not any of the other factors. This allow us to not only consider the properties of the scales independently as in IRT, but also to investigate if there might be overlap between items across and other scales.

In Study 1 for IRT analyses of the polytomous items of the CL scales, we used the Partial Credit Model (PCM; Masters, 1982) in the Digram program (Kreiner & Nielsen, 2013). An overall test of DIF and homogeneity was conducted with Andersen's conditional likelihood test (Andersen, 1973). Item fit was assessed with item rest score correlations (Christensen & Kreiner, 2013). For the analyses of items-wise DIF in relation to gender, age (grouped by 1 = 0-23, and 2 = 23 and above), and course and LD we used Keldermans' likelihood ration test (Kelderman, 1984) and Goodmann and Kruskal's partial gamma correlation (Kreiner & Christensen, 2004).

For Pure PCM models in Study 1 we used Cronbach's Alpha (Cronbach, 1951) to estimate reliability. For scales with evidence of LD we used a Monte Carlo procedure to estimate the reliability since Cronbach's Alpha is prone to inflation for scales with local dependence. To account for false discovery rates due to the multiple testing we used the Benjamini & Hochberg procedure (Benjamini & Hochberg, 1995). E.g., we employed the procedure to test of all possible item pairs in relation to local dependence.

To conduct the CFA we used the Lavaan package (version 0.6-5) in the R statistical programming language (version 3.6.3). To estimate the loading of the model, we used the diagonally least square method (Li, 2016), since the items were ordinal. We used the

Comparative Fit Index (CFI) and the Tucker Lewis Index (TLI) with values above .95 to indicate acceptable fit (Hu & Bentler, 1999). Besides CFI and TLI we used the Root Mean Square Error of Approximation (RMSEA) and the Standardized Root Mean Square Residual (SRMR) were values below .06 and .08, indicate a good fit, respectively (Hu & Bentler, 1999).

Results Study 1

Results of fit to the Partial Credit Model

The Rasch analysis of the IL scale indicated no evidence against the fit to a PCM. The overall test found no evidence of breach of homogeneity, the overall test, or the item-wise tests of DIF in relation to gender, age, or course. There was no evidence against item fit and no evidence of local dependence (see Table 2). The reliability of the scale measured in terms of the Cronbach's Alpha was 0.89. Therefore, we concluded that the scale provided valid and reliable measurement.

The analyses of EL instructions scale exhibited evidence of DIF in relation to course for item 1, such that it was easier for students of psychological testing to endorse the statement *"The instructions and/or explanations during the activity were very unclear"* than for students of educational psychology, despite similar levels of EL related to instructions. When the DIF was added to a graphical log linear Rasch model, then neither the overall test nor the item-wise test showed evidence of DIF. There was no evidence of breach of homogeneity, or against item fit. Finally, there was no evidence of local dependence between items and the reliability was 0.84. Therefore, we concluded that the scale provided valid and reliable measurement.

The Analyses of the Extraneous load scale for noise, showed evidence of both DIF and local dependence. Although, no evidence against item fit and with a reliability coefficient of

0.81, it was not possible to find a working model with LD and DIF which could converge, thus there was no fit to a PCM of a GLLRM for the EL N scale.

For the extraneous load scale in relation to devices, we found evidence of DIF in relation to age for item 2, meaning it was easier for younger students to endorse the statement: *"Messages and notifications from my phone/computer made learning unclear."* and local dependence between item 1 *"My activities on my phone/computer made it difficult to focus on the learning content"*, and item 2 "*Messages and notifications from my phone/computer made learning unclear*". After adding these two deviations from the PCM to a GLLRM, there was no further evidence of DIF or LD, and no evidence against item fit or homogeneity, but the reliability of the scale was only 0.62. Therefore, we conclude that the scale did not fit the PCM, and that we were able to model the DIF and LD, however, the scale had low reliability.

To achieve a working model for the Germane Load scale item 2: *"The activity really enhanced my knowledge and understanding of cognitive load/psychological testing."* was omitted. Further analysis of the remaining three items showed evidence of DIF relative to gender for item 4, Such that it was easier for females to endorse: *"The activity really enhanced my understanding of concepts and definitions"*, despite having the same level of germane load. Furthermore, we found evidence of local dependence between item 3: *"The activity really enhanced my understanding of the theories covered."* and item 4.: *"The activity really enhanced my understanding of concepts and definitions."* After these two instances were added to a GLLRM, there was no further evidence of DIF or LD and no evidence against item fit or against homogeneity, and the scale had a reliability of 0.90. Therefore, we concluded that the scale did not fit a pure Partial Credit Model, but could still provide close to optimal measurement after accounting for DIF and LD.

Results of fit to the Confirmatory Factor Analysis

A Confirmatory Factor analysis was run including all items from all scales with the exception of item 2 from the GL scale because it did not fit the Partial Credit model in the previous analysis , and without the EL Noise scale which did not converge during the PCM analyses. The model grouped each scale's item so they formed a latent construct for each scale, e.g. all IL items loading only on the latent construct of IL. The model achieved acceptable fit values. The CFI was 0.999 and the TLI was 0.999. The RMSEA was < 0.001 and SRMR was .041, thus all values indicated the model was acceptable.

Discussion Study 1

Overall the IRT and CFA analyses provided positive evidence of the construct validity of the MCLS-POL with few minor cases of LD and DIF which could be modeled. However three major issues were identified. The first was that the EL Noise scale would not converge to a meaningful model. Adding instances of local dependence to the model led to other instances other local dependence until the program could no longer converge. Second, although the EL devices scales converged to a model after accounting for LD between two items and accounting for DIF, the scales reliability was lower than conventional cut-off for satisfactory reliability. Finally, item 2 from the GL scale had to be eliminated as it did not fit the model. These issues were dealt with in a revision of the MCLS-POL which is described in Study 2. Overall we found evidence against the validity of the EL noise scale, but we found no evidence against the validity of the other scales.

Study 2

Study 2 was conducted to improve the MCLS-POL based on the results of Study 1. We were interested in investigating the criterion validity of the different sub-scales within the MCLS-POL in addition to testing the reliability and validity of the measure using the PCM and CFA as in Study 1. Sensitivity was tested by using an experimental design where students experienced a lecture in educational psychology either physically or online through Zoom. An experiment was possible because restrictions due to the COVID-19 pandemic meant that approximately half of the students were assigned to a group who had to follow the lecture online instead of physically in order to increase physical distancing in the lecture hall. The students attending the lecture online followed the same lecture as the students who were physically present, while online the students could choose between seeing just the lecture slides or the teacher in front of the lecture slides, while listening to teacher speak. To examine if the uses of scales scores made sense we used the validity frame work of Kane (2013), and examined whether the scales showed meaningful differences such that online students experienced more EL than off-line students as hypothesized in the introduction.

Item revision

Before conducting the study, we reformulated the wording of the items for the EL Noise scale so the item content became more general based on the finding that the scale did not fit the PCM in Study 1. E.g. we changed the wording of item 1 from "*Other students talking in the classroom made it difficult to focus on the learning content*" to "*Noises in the environment made it difficult to focus on the learning content*". (See Table 3 for all items used in Study 2 and Study 3). This was also useful as restrictions due to the COVID-19 pandemic meant that approximately half of the students were assigned to a group who had to follow the lecture online to increase physical distancing in the lecture hall. Given the difference between participating in a lecture

physically and online we expected differences in the EL sub-scales of Noise, and Devices. We also added two more items to the EL Devices sub-scale as the results from Study 1 indicated that the scale had a low reliability.

Sample

Data were collected at a European university during the fall 2020 semester. The psychology students (N = 140) were asked to voluntarily answer a short online survey in relation to their current course in educational psychology. A total of 76.4 % reported being females (n = 107), 22.9 % males (n = 32), and 0.7 % (n = 1) did not wish to answer the question. The mean age was 23.29 with a *SD* of 3.83. Due to the COVID-19 pandemic, the university restricted the number of students who could attend the lecture physically and students were assigned to either attend in person or online through Zoom prior to the lecture. A total of 62 students attended the lecture in person, the rest of the students attended the lecture through the Zoom online streaming service (n = 78). The students experienced the same lecture with the only difference being their presence in the classroom, or experiencing it online through Zoom. The MCLS-POL was administered at the end of the lecture through SurveyMonkey.

Statistical Analyses

In Study 2 for IRT analyses of the polytomous items of the CL scales, we used the Partial Credit Model (PCM; Masters, 1982) in RUMM (Andrich et al., 2003), the switch from Digram to RUMM was made as RUMM is able to handle scales based on only two items which became a necessity in Study 2. An overall test of fit to the PCM was conducted with a chi-square test, where significance indicate misfit in relation to the model (Pallant & Tennant, 2007). Item fit was deemed acceptable if the residuals of the models were within -2.5 and +2.5 (Pallant & Tennant, 2007). Local dependence was assessed by examining the residual correlations between

items, where we expected the residual correlation to be close to zero. We used items residuals above .20 as indicative of local dependence (Christensen et al., 2017). The presence of DIF was examined through analysis of variance in items scores across age, gender, and whether the student was present physically or attended the lecture online, in cases where we tested with multiple items we corrected the *p*-values with the Bonferroni correction to adjust for false discovery rates.

Results for Study 2

Results of fit to the Partial Credit Model

The Rasch analyses of the five scales provide almost no evidence against fit in the overall test or in relation to item fit. A minor deviation was found for the IL scale as the overall test rejected at fit ($p < .001$) however this might be due to item 2 which fit to the PCM was rejected at $p > .05$ but not $p > .01$ after Bonferroni correction. Furthermore the residuals for the item fit was between -2.5 and 2.5, thus we concluded the scale fit.

The only major deviation from the model was related to the EL Devices scale where we identified strong evidence of multidimensionality. After reexamining the wording of the items it was clear that the items were assessing two separate constructs: One measuring the EL from Media with the following items (item 1, "*My activities on my phone/computer made it difficult to focus on the learning content"* and item 2, " *Messages and notifications from my phone/computer made learning unclear"*) and the second measuring EL from devices with the following items (item 4, "*Technical issues made learning ineffective*" and item 5, "*Problems with technology made it difficult to focus*"). Furthermore, item 3 did not fit any of the scales and was eliminated. After the split both scales fit the model and for all scales the reliability was satisfactory (see Table 6).

For the other sub-scales, we only found evidence of one instance of DIF depending on whether the students attended the course physically or online in two items on the EL Instructions scale, such that is was easier for physically present students to endorse the statement in item 2 "*The instructions and/or explanations were, in terms of learning, very ineffective*" than the students attending the lecture online. Opposite of this it was easier for the online students to endorse the statement of item 4 "*Low quality audio made the instructions hard to follow*", than for the physically present students, despite having similar levels of EL in relation to instructions. For the GL scale we again omitted item 2 to achieve a working model, the *p*-value for item fit of item 3 in the GL scale was .0163 (Bonferroni corrected cut-off was .016). however, as the residuals was inside -2.5 to + 2.5 we accepted it. Table 4 illustrates how all other items fit the PCM providing evidence of the validity and reliability of the revised version of the MCLS-POL.

External Validity Results

Table 5 shows the difference between being physically present and attending the lecture online. Independent samples t-tests were conducted to investigate if the differences between the physical and online groups were significant. For the IL scale there was no significant difference between being physically present or attend online ($t_{(138)} = -1.281$, $p = .202$) as expected. Furthermore, EL Noise was significantly higher ($t_{(138)} = -5.795$, $p < .001$) for online students ($M = 2.368$, $SD = 0.962$) than for physically present students ($M = 1.591$, $SD = 0.614$). EL Media was significantly higher ($t_{(138)} = -3.401$, $p < .001$) for online students ($M = 2.820$, $SD = 1.165$) than for physically present students ($M = 2.226$, $SD = 0.904$). EL Devices was significantly higher ($t_{(138)} = -3.290$, $p < .001$) for online student ($M = 2.532$, $SD = 1.070$) than for physically present students ($M = 2.016$, $SD = 0.784$). All of these fit the a-priori hypotheses. However, contrary to the a-priori predictions EL Instructions was significantly higher ($t_{(138)} = -4.558$, $p <$

0.001) for online student (M = 2.234, SD = 0.718) than for physically present students (M =1.762, SD = 0.505). The online students (M = 3.724, SD = 0.713) also experienced significantly ($t_{(138)}$ = 2.071 , p = .040) lower GL than the physically present students (M = 3.944, SD = 0.538). These results suggest that the MCLS-POL is sensitive to differences between students learning in different environments and provides support for the external validity of the measure. However, students also reported different levels of EL related to instructions and GL which was not expected in the a-priori predictions.

Discussion Study 2

Study 2 revealed that the EL Devices scale should be split into two sub-scales in order to create valid measurement. A meaningful categorization was made by creating an EL Media subscale, and an EL Devices subscale. The EL Media sub-scale consisted of the item "*My activities on my phone/computer made it difficult to focus on the learning content*" and the item "*Messages and notifications from my phone/computer made learning unclear*". The EL Devices sub-scale consisted of the item "*Technical issues made learning ineffective*" and the item "*Problems with technology made it difficult to focus*". Due to the item wording (i.e. the first two items pertaining to disturbances from the devices and the second two items pertaining to technology in general) it made sense to split the scale into two distinct scales. Furthermore, comparing the students based on whether they were physically present or attending the lecture online revealed meaningful differences such that the students attending the lecture online experience significantly more EL related to instructions, noise, media ,and devices, as well as significantly less GL than the students who were physically present in the lecture hall. The difference between the groups on IL was not significant. To investigate if these results would replicate in a new setting we conducted a follow-up study.

Study 3

Study 3 was conducted to test the validity of the MCLS-POL in a new context by replicating Study 2 in a sample of psychology master students who were participating in a lecture about psychological testing. The same statistical analyses were conducted as in Study 2.

Sample

Data was collected at a European university during the fall 2020 semester. The psychology master students (N = 119) were asked to voluntarily answer a short online survey in relation to a course in psychological testing. A total of 89.1 % reported being females (n = 106), and 10.9 % males (n = 13). The mean age was 27.25 with an SD of 6.65. Similar to Study 2, students were assigned to attending the course physically or online prior to the lecture but students who were allowed to attend physically were given the option of attending online. A total of 27 students attended the lecture in person, the rest of the students attended the lecture through the Zoom online streaming service (n = 92). The students attending in person had to wear a mask while entering the lecture hall, which they could remove while seated and they had to sit with distance between them. The teacher also had to wear a mask when entering the lecture hall, but the teacher was allowed to remove the mask during the lecture.

Results for Study 3

Results of fit to the Partial Credit Model

The Rasch analyses of the six scales provide almost no evidence against fit in the overall test or in relation to item fit. We only found evidence DIF depending on whether the students attended the course online or by being physically present in relation to two items on the EL Noise scale, such that it was easier for physically present students to endorse "*Noises in the environment made it difficult to focus on the learning content*", while it was easier for student

attending online to endorse "*Distractions in the environment made learning ineffective*", despite having the same level of EL related to noise. Again we split the EL device scale into two separate two-item scales. One measuring the EL from media and the second measuring EL from devices. After the split both scales fit the model and for all scales the reliability was above .80 and thus satisfactory (see Table 6).

External Validity Results

Table 7 shows the difference between being physically present and attending the lecture online. The EL noise and the EL devices scales showed significant differences across type of attendance as predicted. For the EL Noises scale scales the students who attended the lecture online (M = 2.370, SD = 0.745) experienced significantly ($t_{(117)}$ = -3.878 , $p < .001$) more EL related to noises than the students who were physically present (M = 1.741, SD = 0.669). Similarly, on the EL Devices scale the students who attended the lecture online (M = 2.179, SD = 1.015) experienced significantly ($t_{(117)}$ = -2.076 , $p = .040$) more EL related to devices than the students who were physically present (M = 1.740, SD = 0.764). However, although the online lecture group (M = 2.696, SD = 1.021) also experienced more EL related to media than the students who were physically present (M = 2.333, SD = 1.047) this difference did not reach statistical significance ($t_{(117)}$ = -1.608 , $p = .111$). Finally, the difference between the groups on IL, EL related to instructions, and GL were not statistically significant as predicted. The results suggest that the EL noise and EL devices scales within the MCLS-POL is sensitive to differences between students learning in different environments and provides support for the external validity of the measure.

Discussion Study 3

Study 3 showed that all six scales provide valid measurement. Furthermore, comparing the students based on whether they were physically present or attending the lecture online revealed differences in a way that the students attending the lecture online experience significantly more EL related to noises and devices than the students who were physically present, other than those two scales there were no significant difference in the experienced CL. These results suggest that it is important to have a multidimensional conceptualization of EL as different components of EL can influence learning in physical and online environments differently.

It is not immediately clear why the differences in mean between students attending the lecture physically and students attending online in Study 3 are not similar to that of students in Study 2. One explanation might be that the master students attending Psychological Testing were more accustomed to lectures than the bachelor students attending Educational Psychology. Thus, we might reason that more experienced learners are less hampered by different types of extraneous load despite attending lectures online

Results of Combining Data from Study 2 and Study 3

Results of the Confirmatory Factor Analysis

Before comparing the sum scores of Study 2 and Study 3 we conducted a confirmatory factor analysis with all items loading on their respective factors by combining data from Study 2 and Study 3. The fit indices were CFI = .99, TLI = .99, and both RMSEA and SRMR = .06 which are all satisfactory for the six factor model.

Comparing Cognitive Load across Study 2 and Study 3

Although the samples of psychology students in Study 2 and Study 3 differed because the students in the educational psychology course were second year bachelor students and the

students in the psychological testing course were master students, we were interested in comparing the cognitive load ratings across the studies. Since psychological testing is considered a more cognitive straining course by many students, we wanted to compare the scales across the two courses. To ensure the scales were comparable across the studies we performed a CFA which yielded satisfactory fit indices. When comparing across courses we found IL to be significantly ($t_{(257)} = -5.317$, $p < .001$) higher for students from psychological testing ($M = 3.18$, $SD = 0.81$) than for students of educational psychology($M = 2.67$, $SD = 0.72$). Similarly, We found EL instruction to be significantly ($t_{(257)} = -5.625$, $p < .001$) higher for students from psychological testing ($M = 2.53$, $SD = 0.76$) than for students of educational psychology ($M = 2.03$, $SD = 0.67$). On the other hand GL was significantly ($t_{(257)} = 7.158$, $p < .001$) lower for students of psychological testing($M = 3.21$, $SD = 0.65$) than for students of educational psychology ($M = 3.82$, $SD = 0.65$, See Table 8). The differences between the groups were not significant for the other scales.

Conclusion

Through three studies we describe the further development and validation of the MCLS-POL. We provide evidence of the validity and reliability of the expanded CLS which supports the multidimensional conceptualization of cognitive load. Overall there was evidence of meaningful external validity in terms of meaningful group difference between students experiencing a lecture physically and students who experience the same lecture online. However, since one of the scale was split into two separate scale with only two items each, we highly recommend that researchers wishing to use these scale enhance these two scales by adding more items to them.

The studies also reveals meaningful challenges that students as well as lectures face as more teaching is conducted online. The results from Study 2 revealed that the students attending the lectures online experienced more EL on all four EL load sub-scales, meaning that they experienced extraneous cognitive load related to instructions, noises, media, and devices, which was greater than what the students who were physically present experienced. Furthermore, the physically present students in Study 2 also reported higher GL than the students who attended the lecture online.

The students attending the lecture online in Study 3 similarly experienced more cognitive load related to EL from media and devices than the physically present students. The finding that there were no differences on IL between the students who experienced the lecture physically or online in Studies 2 and 3, but there were differences in different components of EL suggests that the MCLS-POL is sensitive at identifying different components of cognitive load.

A limitation of these studies is the lack of measurement of learning, and the subsequent hypothetical analyses of differences in learning across the students attending either online or off-line and correlations between the CL scales and learning outcome. However, as we examined the cognitive load across differing course, creating a measure of learning with similar properties across courses was difficult. Future studies might address this by examining learning in just one type of course.

The reason for providing online lectures in Studies 2 and 3 was due to the extraordinary consequences of the COVID-19 pandemic in 2020. However, many universities are aiming at providing more online lectures, due to several advantages such as accessibility issues e.g. the ability to reach more students and allow students to access high quality educational opportunities even though they are unable to be physically present (Cascaval et al., 2008; French & Kennedy,

2017; Makransky, Mayer, et al., 2019; Waschull, 2001). In contrast to the benefits this article highlights some of the caveats of online teaching related to the strain it can provide for students in terms of more EL, which should be addressed when conducting online lectures.

Data from all studies will be made available upon reasonable request

References

Ali, Sayed Abas Ai. 2013. “Study effects of school noise on learning achievement and annoyance in Assiut City, Egypt.” Applied Acoustics 74(4):602–6. doi: 10.1016/j.apacoust.2012.10.011.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123–140. https://doi.org/10.1007/BF02291180

Andersen, M. S., & Makransky, G. (2020). The validation and further development of a multidimensional cognitive load scale for virtual environments. *Journal of Computer Assisted Learning*, *n/a*(n/a). https://doi.org/10.1111/jcal.12478

Andrich, D., Sheridan, B., & Luo, G. (2003). *RUMM2020: Rasch Unidimensional Measurement Models [Computer Software]*.

Antonenko, P., Paas, F., Grabner, R., & van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review*, *22*(4), 425–438. https://doi.org/10.1007/s10648-010-9130-y

Ayres, P. (2006). Impact of reducing intrinsic cognitive load on learning in a mathematical domain. *Applied Cognitive Psychology*, *20*(3), 287–298. https://doi.org/10.1002/acp.1245

Ayres, P. (2018). Subjective measures of Cognitive Load—What can they reliably measure? In *Cognitive Load Measurement and Application—A Theoretical Framework for Meaningful Research and Practice* (1st ed.).

Baceviciute, S., Mottelson, A., Terkildsen, T., & Makransky, G. (2020). Investigating representation of text and audio in educational VR using learning outcomes and EEG. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376872

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Blasiman, R. N., Larabee, D., & Fabry, D. (2018). Distracted students: A comparison of multiple types of distractions on learning in online lectures. *Scholarship of Teaching and Learning in Psychology*, *4*(4), 222–230. https://doi.org/10.1037/stl0000122

Cascaval, R. C., Fogler, K. A., Abrams, G. D., & Durham, R. L. (2008). Evaluating the benefits of providing archived online lectures to in-class math students. *Journal of Asynchronous Learning Networks*, *12*, 61–70.

Cerdan, R., Candel, C., & Leppink, J. (2018). Cognitive load and learning in the study of multiple documents. *Frontiers in Education*, *3*. https://doi.org/10.3389/feduc.2018.00059

Chen, Q., & Yan, Z. (2016). Does multitasking with mobile phones affect learning? A review. *Computers in Human Behavior*, *54*, 34–42. https://doi.org/10.1016/j.chb.2015.07.047

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. ternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1, pp. 7–76). Lawrence Erlbaum Associates, Inc.

Christensen, K. B., & Kreiner, S. (2013). Item fit statistics. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch Models in health* (Vol. 2013, pp. 83–104). ISTE and John Wiley & Sons, Inc. doi:10.1002/ 9781118574454.ch5

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Cierniak, G., Gerjets, P., & Scheiter, K. (2009). Expertise reversal in multimedia learning: Subjective load ratings and viewing behavior as cognitive process indicators. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *31*(31). https://escholarship.org/uc/item/99r5z5fb

Costley, J., Fanguy, M., Lange, C., & Baldwin, M. (2020). The effects of video lecture viewing strategies on cognitive load. *Journal of Computing in Higher Education*. https://doi.org/10.1007/s12528-020-09254-y

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Elkaseh, A., Wong, K. W., & Fung, C. C. (2015). A review of the critical success factors of implementing e-learning in higher education. *The International Journal of Technologies in Learning*, *21*(2), 1–13. https://doi.org/10.18848/2327-0144/CGP/v22i02/49160

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists* (pp. xi, 371). Lawrence Erlbaum Associates Publishers.

French, S., & Kennedy, G. (2017). Reassessing the value of university lectures. *Teaching in Higher Education*, *22*(6), 639–654. https://doi.org/10.1080/13562517.2016.1273213

Hadie, S. N. H., & Yusoff, M. S. B. (2016). Assessing the validity of the cognitive load scale in a problem-based learning setting. *Journal of Taibah University Medical Sciences*, *11*(3), 194–202. https://doi.org/10.1016/j.jtumed.2016.04.001

Hamon, A., & Mesbah, M. (2002). Questionaires reliability under the Rasch model. In M. Mesbah, B. F. Cole, & M.-L. T. Lee, *Statistical Methods for Quality of Life Studies: Design, Measurements and Analysis* (2002 edition). Kluwer Academic Publishers.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

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

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kappe, R., & van der Flier, H. (2012). Predicting academic success in higher education: What's more important than being smart? *European Journal of Psychology of Education*, *27*(4), 605–619. https://doi.org/10.1007/s10212-011-0099-9

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika*, *49*(2), 223–245. https://doi.org/10.1007/BF02294174

Kirschner, P. A., Ayres, P., & Chandler, P. (2011). Contemporary cognitive load theory research: The good, the bad and the ugly. *Computers in Human Behavior*, *27*(1), 99–105. https://doi.org/10.1016/j.chb.2010.06.025

Klatte, M., Bergstroem, K., & Lachmann, T. (2013). Does noise affect learning? A short review on noise effects on cognitive performance in children. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00578

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01997

Klepsch, M., & Seufert, T. (2021). Making an effort versus experiencing load. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.645284

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. Guilford Publications.

Kohnke, L., & Moorhouse, B. L. (2020). Facilitating synchronous online language learning through zoom. *RELC Journal*, 0033688220937235. https://doi.org/10.1177/0033688220937235

König, J., Jäger-Biela, D. J., & Glutsch, N. (2020). Adapting to online teaching during COVID-19 school closure: Teacher education and teacher competence effects among early career teachers in Germany. *European Journal of Teacher Education*, *43*(4), 608–622. https://doi.org/10.1080/02619768.2020.1809650

Kreiner, S., & Christensen, K. B. (2004). Analysis of local dependence and multidimensionality in graphical loglinear Rasch models. *Communications in Statistics - Theory and Methods*, *33*(6), 1239–1276. https://doi.org/10.1081/STA-120030148

Kreiner, S., & Nielsen, T. (2013). *Item analysis in DIGRAM 3.04: Part I: Guided tours* [Research Report]. Department of Biostastistics, University of Copenhagen.

Kuznekoff, J. H., & Titsworth, S. (2013). The impact of mobile phone usage on student learning. *Communication Education*, *62*(3), 233–252. https://doi.org/10.1080/03634523.2013.767917

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, *45*(4), 1058–1072. https://doi.org/10.3758/s13428-013-0334-1

Li, C.-H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, *21*(3), 369–387. https://doi.org/10.1037/met0000093

Makransky, G., Mayer, R. E., Veitch, N., Hood, M., Christensen, K. B., & Gadegaard, H. (2019). Equivalence of using a desktop virtual reality science simulation at home and in class. *PLOS ONE*, *14*(4), e0214944. https://doi.org/10.1371/journal.pone.0214944

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Role of subjective and objective measures of cognitive processing during learning in explaining the spatial contiguity effect. *Learning and Instruction*, *61*, 23–34. https://doi.org/10.1016/j.learninstruc.2018.12.001

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

McKenzie, K., & Schweitzer, R. (2001). Who succeeds at university? Factors predicting academic performance in first year Australian university students. *Higher Education Research & Development*, *20*(1), 21–33. https://doi.org/10.1080/07924360120043621

Minkley, N., Xu, K., & Krell, M. (2021). Analyzing relationships between causal and assessment factors of cognitive load: Associations between objective and subjective measures of cognitive load, stress, interest, and self-concept. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.632907

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *The British Journal of Clinical Psychology*, *46*(Pt 1), 1–18. https://doi.org/10.1348/014466506x96931

Rosenbaum, P. R. (1989). Criterion-related construct validity. *Psychometrika*, *54*(4), 625–633. https://doi.org/10.1007/BF02296400

Scharinger, C., Schüler, A., & Gerjets, P. (2020). Using eye-tracking and EEG to study the mental processing demands during learning of text-picture combinations. *International Journal of Psychophysiology*, *158*, 201–214. https://doi.org/10.1016/j.ijpsycho.2020.09.014

Servilha, E. A. M., Delatti, M. de A., Servilha, E. A. M., & Delatti, M. de A. (2014). College students' perception of classroom noise and its consequences on learning quality. *Audiology - Communication Research*, *19*(2), 138–144. https://doi.org/10.1590/S2317-64312014000200007

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*(2), 123–138. https://doi.org/10.1007/s10648-010-9128-5

Sweller, J., Ayres, P., & Kalyuga, S. (2011a). Categories of knowledge: An evolutionary approach. In J. Sweller, P. Ayres, & S. Kalyuga (Eds.), *Cognitive Load Theory* (pp. 3–14). Springer New York. https://doi.org/10.1007/978-1-4419-8126-4_1

Sweller, J., Ayres, P., & Kalyuga, S. (2011b). Intrinsic and extraneous cognitive load. In J. Sweller, P. Ayres, & S. Kalyuga (Eds.), *Cognitive Load Theory* (pp. 57–69). Springer New York. https://doi.org/10.1007/978-1-4419-8126-4_5

Sweller, J., Ayres, P., & Kalyuga, S. (2011c). Measuring cognitive load. In J. Sweller, P. Ayres, & S. Kalyuga (Eds.), *Cognitive Load Theory* (pp. 71–85). Springer New York. https://doi.org/10.1007/978-1-4419-8126-4_6

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. https://doi.org/10.1023/A:1022193728205

Tindall-Ford, S., Agostinho, S., & Sweller, J. (Eds.). (2019). *Advances in cognitive load theory: Rethinking teaching*.

Waschull, S. B. (2001). The online delivery of psychology courses: Attrition, performance, and evaluation. *Teaching of Psychology*, *28*(2), 143–147. https://doi.org/10.1207/S15328023TOP2802_15

Zheng, R., & Cook, A. (2012). Solving complex problems: A convergent approach to cognitive load measurement. *British Journal of Educational Technology*, *43*(2), 233–246. https://doi.org/10.1111/j.1467-8535.2010.01169.x

Zukić, M., Đapo, N., & Husremović, D. (2016). Construct and predictive validity of an instrument for measuring intrinsic, extraneous and germane cognitive load. *Universal Journal of Psychology*, *4*(5), 242–248. https://doi.org/10.13189/ujp.2016.040505

Zureick, A. H., Burk-Rafel, J., Purkiss, J. A., & Hortsch, M. (2018). The interrupted learner: How distractions during live and video lectures influence learning outcomes. *Anatomical Sciences Education*, *11*(4), 366–376. https://doi.org/10.1002/ase.1754

Table 1: Items and scales included in the Study 1

Scale	
IL	The topics covered in the activity were very complex.
IL	The activity covered theories that I perceived as very complex.
IL	The activity covered concepts and definitions that I perceived as very complex.
EL ins	The instructions and/or explanations during the activity were very unclear.
EL ins	The instructions and/or explanations were, in terms of learning, very ineffective.
El ins	The instructions and/or explanations were full of unclear language.
EL noi	Other students talking in the classroom made it difficult to focus on the learning content.
EL noi	Students talking to me during the activity made learning ineffective.
EL noi	Other noises and distractions during the activity made it hard to learn.
EL dev	My activities on my phone/computer made it difficult to focus on the learning content.
EL dev	Messages and notifications from my phone/computer made learning unclear.
EL dev	Others' phone/computer use distracted me, making it hard to learn.
GL	The activity really enhanced my understanding of the topic(s) covered.
GL	The activity really enhanced my knowledge and understanding of cognitive load.
GL	The activity really enhanced my understanding of the theories covered.
GL	The activity really enhanced my understanding of concepts and definitions.

IL = Intrinsic Load, EL Ins = Extraneous Load Instructions, EL Noi = Extraneous Load Noises, EL Dev = Extraneous Load Devices, GL = Germane Load.

Table 2: Results for the Rasch analyses of the scales in Study 1

Scale	Overall test	Item fit	DIF gender	DIF age	DIF course	LD	*r*
IL	✓	✓	✓	✓	✓	✓	.89
EL Ins	✓	✓	✓	✓	%	✓	.84
EL Noi	%	✓	%	%	%	%	.81
EL Dev	✓	✓	✓	%	✓	%	.62
GL	✓	✓	%	✓	✓	%[6]	.90

Note. IL = Intrinsic Load, EL Ins = Extraneous Load Instructions, EL Noi = Extraneous Load Noises, EL Dev = Extraneous Load Devices, GL = Germane Load,
DIF = Differential Item Functioning, LD = Local Dependence, r = reliability.

Table 3: Items and scales included in the Study 2 and Study 3

Scale	
IL	The topics covered in the activity were very complex.
IL	The activity covered theories that I perceived as very complex.
IL	The activity covered concepts and definitions that I perceived as very complex.
EL Ins	The instructions and/or explanations during the activity were very unclear.
EL Ins	The instructions and/or explanations were, in terms of learning, very ineffective.
EL Ins	The instructions and/or explanations were full of unclear language.
EL Ins	Low quality audio made the instructions hard to follow.
EL Noi	Noises in the environment made it difficult to focus on the learning content.
EL Noi	Distractions in the environment made learning ineffective.
EL Noi	Unrelated events occurring in the environment made it difficult to focus.
EL Dev[1]	My activities on my phone/computer made it difficult to focus on the learning content.
EL Dev[1]	Messages and notifications from my phone/computer made learning unclear.
EL Dev[2]	Others' phone/computer use distracted me, making it hard to learn.
EL Dev	Technical issues made learning ineffective.
EL Dev	Problems with technology made it difficult to focus.
GL	The activity really enhanced my understanding of the topic(s) covered.
GL	The activity really enhanced my knowledge and understanding of [course subject].
GL	The activity really enhanced my understanding of the theories covered.
GL	The activity really enhanced my understanding of concepts and definitions.

IL = Intrinsic Load, EL Ins = Extraneous Load Instructions, EL Noi = Extraneous Load Noises, EL Dev = Extraneous Load Devices, GL = Germane Load. 1 = was combined to make a EL Media scale, 2 = was omitted from final analyses.

Table 4: Results for the Rasch analyses of the scales in Study 2 in RUMM

Scale	Overall test	Item fit	DIF gender	DIF age	DIF location	LD	*r*
IL	%	✓	✓	✓	✓	✓	.86
EL Ins	✓	✓	✓	✓	%	✓	.73
EL Noi	✓	✓	✓	✓	✓	✓	.85
EL Med	✓	✓	✓	✓	✓	✓	.85
EL Dev	✓	✓	✓	✓	✓	✓	.87
GL	✓	✓	✓	✓	✓	✓	.88

Note. IL = Intrinsic Load, EL Ins = Extraneous Load Instructions, EL Noi = Extraneous Load noise, EL Med = Extraneous Load Media, EL Dev = Extraneous Load devices, GL = Germane Load, DIF = Differential Item Functioning, LD = Local Dependence, r = reliability.

Table 5: A comparison of the students who attended the course online and those who were physically present on the scales in the MCLS-POL in Study 2.

Scale	Online		Physical present		*t*(138)	*p*	Cohen' s *d*
	M	*SD*	*M*	*SD*			
IL	2.773	0.717	2.586	0.731	-1.281	.202	0.259
EL Ins	**2.234**	**0.718**	**1.762**	**0.505**	**-4.558**	**<.001**	**0.746**
EL Noi	**2.368**	**0.962**	**1.591**	**0.614**	**-5.795**	**<.001**	**0.940**
EL Med	**2.820**	**1.165**	**2.226**	**0.904**	**-3.401**	**<.001**	**0.746**
EL Dev	**2.532**	**1.070**	**2.016**	**0.784**	**-3.290**	**<.001**	**0.541**
GL	**3.724**	**0.713**	**3.944**	**0.538**	**2.071**	**.040**	**0.343**

Note. IL = Intrinsic Load, EL Ins = Extraneous Load instructions, EL Noi = Extraneous Load Noises, EL Med = Extraneous Load Media, EL Dev = Extraneous Load Devices, GL = Germane.

Table 6: Results for the Rasch analyses of the scales in Study 3 RUMM

Scale	Overall test	Item fit	DIF gender	DIF age	DIF Zoom	LD	*R*
IL	✓	✓	✓	✓	✓	✓	.93
EL Ins[2]	✓	✓	✓	✓	✓	✓	.84
EL Noi	✓	✓	✓	✓	%	✓	.81
EL Med	✓	✓	✓	✓	✓	✓	.85
EL Dev	✓	✓	✓	✓	✓	✓	.90
GL[7]	✓	✓	✓	✓	✓	✓	.90

Note. IL = Intrinsic Load, EL Ins = Extraneous Load Instructions, EL Noi = Extraneous Load noise, EL Med = Extraneous Load Media EL Dev = Extraneous Load Devices, GL = Germane Load, DIF = Differential Item Functioning, LD = Local Dependence, r = reliability.

Table 7: A comparison of the students who attended the course online and those who were physically present on the scales in the MCLS-POL in Study 3.

Scale	Online		Physical present		*t*(117)	*p*	Cohen' s *d*
	M	*SD*	*M*	*SD*			
IL	3.181	0.836	3.200	0.718	0.092	.927	0.023
EL Ins	2.544	0.745	2.463	0.814	-0.848	.629	0.106
EL Noi	**2.370**	**0.950**	**1.741**	**0.669**	**-3.878**	**<.001**	**0.703**
EL Med	2.696	1.021	2.333	1.047	-1.608	.111	0.354
EL Dev	**2.179**	**1.015**	**1.740**	**0.764**	**-2.076**	**.040**	**0.455**
GL	3.266	0.677	3.037	0.822	-1.471	.144	0.322

Note. IL = Intrinsic Load, EL Ins = Extraneous Load Instructions, EL Noi = Extraneous Load Noises, EL Med = Extraneous Load Media, EL Dev = Extraneous Load Devices, GL = Germane Load.

Table 8: Difference between scores on the MCLS-POL in Study 2 and Study 3

Scale/course	Educational Psychology		Psychological Testing		*t*	*p*	Cohen' s *d*
	M	*SD*	*M*	*SD*			
IL	2.67	0.72	3.18	0.81	-5.317	<.001	0.67
EL Ins	2.03	0.67	2.53	0.76	-5.625	<.001	0.70
EL Noi	2.02	0.91	2.23	0.93	-1.775	.077	0.23
EL Med	2.56	1.09	2.61	1.04	-0.423	.673	0.05
EL Dev	2.30	0.98	2.08	0.98	1.827	.069	0.22
GL	3.82	0.65	3.21	0.72	7.158	<.001	0.89

Note. IL = Intrinsic Load, EL Ins = Extraneous Load Instructions, EL Noi = Extraneous Load Noises, EL Med = Extraneous Load Media, EL Dev = Extraneous Load Devices, GL = Germane Load.