The Validation and Further Development of a Multidimensional Cognitive Load Scale for Virtual Environments

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

The Validation and Further Development of a Multidimensional Cognitive Load Scale for Virtual Environments

Abstract

Measuring cognitive load is important in virtual learning environments (VLE). Thus, valid and reliable measures of cognitive load are important to support instructional design in VLE. Through three studies, we investigated the validity and reliability of Leppink's Cognitive Load Scale (CLS) and developed the extraneous cognitive load (EL) dimension into three sub-scales relevant for VLE: EL instructions, EL interaction, and EL environment. We investigated the validity of the measures using the Partial Credit Model (PCM), Confirmatory Factor Analysis (CFA), and correlations with retention tests. Study 1 (n = 73) investigated the adapted version of the CLS. Study 2 describes the development and validation of the Multidimensional Cognitive Load Scale for Virtual Environments (MCLSVE), with 140 students in higher education. Study 3 tested the generalizability of the results with 121 higher education students in a more complicated VLE. The results provide initial evidence for the validity and reliability of the MCLSVE.

Keywords: cognitive load, virtual reality, confirmatory factor analysis, item response theory, virtual simulations

Introduction

Cognitive Load Theory (CLT; Sweller et al., 2011b) and the Cognitive Theory of Multimedia Learning (CTML; Mayer, 2009) have been central theoretical frameworks for identifying a number of factors that either facilitate or inhibit learning. The fundamental basis for the theories is that working memory (WM) is limited in terms of capacity and duration, meaning that people can only process a limited amount of information in their WM and that they are only able to retain it for a short period of time (Baddeley, 1990). In contrast, long-term memory (LTM) capacity is greater because information is stored in structures called schema (Chi et al., 1982). Before encoding it in LTM, the mind processes and integrates new information in WM (Sweller et al., 2011b). The processing that occurs in WM during learning creates cognitive load. A basic assumption of CLT is that cognitive overload can occur when the complexity of a task exceeds the capacity of the WM. The overload inhibits learning due to the limited capacity of WM.

Although CLT was originally developed based on the idea that cognitive load was a unidimensional construct consisting of total working memory (Ayres, 2018), empirical studies have caused the theory to evolve. Today, cognitive load is generally described as consisting of: intrinsic, extraneous, and germane load (Sweller et al., 2011a). Intrinsic cognitive load is related to how complex the learner finds the learning material (Sweller et al., 1998) and it varies based on the nature of the task. Extraneous cognitive load is not intrinsic to the learning material and is instead caused by processing unnecessary information, which is created by instructional design decisions, i.e. including non-essential information in learning material which means students have to expend cognitive resources on filtering out the non-essential material (Sweller et al., 2011a). Finally, germane cognitive load is related to the germane resources devoted to information relevant for learning (Ayres, 2018). CLT and the CTML describe three kinds of instructional design goals that include managing essential processing related to the intrinsic nature of the learning material; reducing extraneous processing that is not necessary for learning essential material; and fostering generative processing thereby facilitating learners to exert effort during learning (Mayer & Moreno, 2003).

Given the influence of CLT for instructional design and the understanding of learning, an important topic of research has been the measurement of cognitive load. The most common way to measure cognitive load is through subjective measures. Previously, the most commonly used subjective measure of CL is a single item developed by Paas (1992; Klepsch et al., 2017; Krell,

2017) which measures mental effort using a nine-point Likert scale ranging from *very very low mental effort* (1) to *very very high mental effort* (9). A single item measure of the perceived difficulty of the learning material using the same format has also been used (Ayres, 2006; Cierniak, Gerjets, et al., 2009; Cierniak, Scheiter, et al., 2009; Paas, 1992). Furthermore, by aligning the wording of subjective items to measure different types of load, Cierniak, Scheiter, and Gerjets (2009, p. 318) use the same scale as in Paas (1992) to measure different components of CL. The new wordings are "How difficult was the learning content for you?" for intrinsic cognitive load, and, "How difficult was it for you to learn with the material?" for extraneous cognitive load. Although single item scales are quick and easy to use, they often fall short when the objective is to differentiate between different types of cognitive load. For instance, the simplicity of the items makes it difficult for learners to make distinctions between having trouble due to the complexity of material or inadequate instructional design (Kirschner et al., 2011).

A lack of a multidimensional psychometrically validated scale to measure cognitive load led Leppink and colleagues (2013) to use factor analysis to develop a 10-item instrument that measures the three dimensions of cognitive load. The Cognitive Load Scale (CLS) refers to the context of learning statistics and includes three items to measure intrinsic load by having learners rate the overall complexity as well as more specific components such as formulas, concepts, and definitions of a learning activity. Furthermore, they measured Extraneous Load (EL) using three items that refer to the clarity and efficiency of the instructions and/or explanations used in the learning activity. Finally, Germane Load (GL) is measured with four items that require the learner to rate how well the activity enhanced their understanding of factors such as topic, knowledge of statistics, formulas, concepts, and definitions in the learning activity. Other studies have also developed and validated multidimensional scales designed to assess cognitive load. Klepsch and colleagues (2017) developed and validated a scales measuring IL, EL and GL. Krell and colleagues (2017) developed a rating scale instrument to measure mental load and mental effort. These measures operationalize CL related to given tasks, whereas the CLS developed by Leppink and colleagues operationalizes CL related to an activity, which makes it more relevant for the purpose of the current study. Furthermore, the CLS is currently the most widely used and researched multidimensional scale to measure cognitive load.

In this article, we first aim to validate the CLS using Item Response Theory (IRT), Confirmatory Factor Analysis (CFA), and by investigating correlations with other measures in

Study 1. Then, in Study 2, we describe the development and validation of the Multidimensional Cognitive Load Scale for Virtual Environments (MCLSVE). The MCLSVE is an adapted and expanded version of the CLS, which includes three sub-scales of extraneous cognitive load that are relevant when learning with immersive virtual reality (IVR). We further investigate the validity of the sub-scales using IRT and CFA, explore their predictive validity, and examine if the scales can differentiate between easy or difficult learning materials in Studies 2 and 3.

Literature Review

Using Confirmatory Factor Analysis (CFA; Kline, 2011) to assess the construct validity of the CLS, Leppink and colleagues reported acceptable fit indices for a three factor model (Leppink et al., 2013). The three factor model was further supported by CFA in a study of medical students, which rejected a one factor model (Hadie & Yusoff, 2016). Reliability was reported as Cronbach's alpha from .75-.82 by Leppink (2013) and ranging from .82-.95 by Hadie and Yusoff (2016). A series of studies (Leppink et al., 2013, 2014) have further developed and validated the scales in different contexts using exploratory and confirmatory factor analysis. Although the measure by Leppink and colleagues (Leppink et al., 2013, 2014) provides learning science researchers with a potential tool for measuring cognitive load as a multidimensional construct, several challenges still exist. Firstly, although initial evidence of the reliability and validity of the scales exist, more research investigating the validity of the items in more detail using modern psychometric approaches such as Item Response Theory (IRT) would strengthen an understanding of the psychometric properties and limitations of the scales. Secondly, more research is needed to investigate if the scales predict objective criteria such as learning outcomes in different settings, which could provide additional evidence of the external validity of the scales across settings.

Several methods for validating questionnaires exist, for example, methods to investigate internal validity such as IRT (Hambleton, 1991), or CFA (Kline, 2011), and methods for investigating external criterion validity such as the correlation with external measures. IRT and CFA have different merits, but complement each other well as they are both latent variable models and together they can provide information about the structure of the instrument and detailed information about how well specific items function (Makransky, Lilleholt & Aaby, 2017).

IRT allows for investigating the validity of items within each sub-dimension based on a number of assumptions thereby providing details about the properties of each item within each sub-dimension independently. Rasch models (Rasch, 1961) are a specific group of IRT models that are relevant when total scores are used, and the analyses allow for the investigation of the validity of items based on several assumptions including unidimensionality, monotonicity, local independence, measurement invariance, and homogeneity, which can provide valuable information regarding scale validation and further development of items.

CFA models are optimal for investigating the construct validity of a measure when there is an a-priori theoretical or empirical hypothesis about its structure. For instance, CFA is ideal for investigating if the 10 items in the CLS (Leppink et al., 2013) actually measure the three dimensional description of CL including intrinsic, extraneous, and germane CL proposed by Sweller and colleagues (1998).

A third challenge with the measurement of cognitive load using existing measures such as the CLS (Leppink et al., 2013; 2014), from an instructional design perspective, is that the operationalization of the extraneous cognitive load dimension is specifically related to the instructions and/or explanations of an activity thereby ignoring other factors that could lead to extraneous cognitive load. A multidimensional understanding of extraneous cognitive load is becoming increasingly important with the increase in technology enhanced learning applications and the widespread use of virtual learning tools because factors other than instructions and/or explanations can increase cognitive load (Makransky, Terkildsen & Mayer, 2019b). For instance, a number of studies have identified cognitive load as a specifically important component of understanding the learning process when learning in IVR. The studies describe other factors such as the complexity of virtual learning environments and the interaction features within these environments as important components of extraneous cognitive load that can impair learning (Makransky, et al., 2019b; Meyer et al., 2019; Parong & Mayer, 2018).

These articles suggest that learning in IVR leads to higher cognitive load than when learning in less immersive media and propose that it is essential that instructional designers consider cognitive load when designing IVR learning tools. Thus, the operationalization of extraneous cognitive load as a product of instructions and/or explanations is not sufficient for learning in immersive environments where components such as the interaction, environment, and instructions all can lead to differential levels of extraneous processing that impedes learning. It is

therefore important to differentiate these factors in order to identify how to optimize instructional design in immersive media, and in order to make meaningful comparisons with less immersive media. For instance, Makransky and colleagues (2019b) compared learning outcomes from a science simulation presented in either immersive or desktop VR and found that the IVR science simulation lead to more presence but less learning than the same simulation on a desktop computer. They also measured cognitive load using electroencephalogram (EEG) and explained their results based on higher levels of extraneous cognitive load in the IVR condition, but also acknowledged that it was not possible to differentiate between different potentially important sources of cognitive load. They propose that the extraneous cognitive load in IVR could have been a consequence of the novelty in the control device in IVR vs. the familiarity of using a mouse in the desktop simulation. However, they also suggest that extraneous CL could be a consequence of having to deal with a complex 3D virtual environment which has previously been referred to as "bells and whistles or a seductive detail" by Moreno and Mayer (2002) when the environment is not directly relevant for learning.

In summary, Leppink and colleagues have developed a validated questionnaire that has been widely used since its inception but more research is needed to further validate the existing instrument, and there is a need to further develop the extraneous cognitive load dimension to include several sub-dimensions that are becoming increasingly relevant when learning in virtual learning environments.

Methods

Item Response Theory, Confirmatory Factor Analysis, and External Validity

IRT and CFA were applied to investigate the psychometric properties of an adapted version of the CLS (Study 1) and the Multidimensional CLS for Virtual Environments (MCLSVE; Study 2 and 3). IRT's strength is that it provides highly detailed analyses of each item in relation to its intended sub-scale, such that a probability function for endorsing each item dependent on the level of trait that is being measured is estimated. Using IRT, the psychometric properties of the items and sub-scales are assessed in relation to five assumptions (Rosenbaum, 1989) of the data: (a) the unidimensionality of each sub-scale (e.g., that the set of items that purportedly measuring intrinsic load all do measure the same construct). (b) The monotonicity of the item (e.g., the probability of endorsing the item increases, as a student's level of the trait of intrinsic load increases). (c) Local independence of the items (i.e., the items are conditionally

independent after accounting for the latent trait). (d) The items are measurement invariant, (i.e., they exhibit no Differential Item Functioning; DIF) meaning that all persons with the same level of for instance intrinsic load have the same probability of endorsing an item regardless of gender or age. (e) The items are homogeneous (e.g., the rank order of the items is constant regardless of levels of intrinsic load). The fifth requirement of homogeneity is unique for the Rasch model family, which is used in this study.

CFA provides the opportunity to model several dimensions in relation to each other. Thus, after the individual properties of each sub-scale were established using IRT, CFA was used to test the fit of data to a multidimensional model including all sub-scales simultaneously. That is, the model contained all cognitive load sub-scales and the fit of this total model was evaluated as opposed to investigating each sub-scale separately as in IRT.

The IRT analyses were conducted using Digram (Kreiner & Nielsen, 2013) which employs the Partial Credit model (PCM; Masters, 1982) for polytomous items. Andersen's conditional likelihood-ratio test (Andersen, 1973) was used to test for overall homogeneity and overall DIF. Individual item fit and thus also monotonicity was investigated using observed and expected item rest-score correlations (Christensen & Kreiner, 2013). Keldermans's likelihood-ratio test (Kelderman, 1984) was used to test for local dependence and DIF, and the strength of the correlation was assessed with Goodmann and Kruskal partial gamma correlation (Kreiner & Christensen, 2004).

Reliability was estimated using Cronbach's alpha (Cronbach, 1951) for the pure Rasch model. Since Cronbach's alpha is inflated when local dependence is present, a Monte Carlo procedure (Hamon & Mesbah, 2002) was used for scales with local dependence as the procedure accounts for local dependence. False discovery (a spurious significance test, due to multiple testing) rates was accounted for by the Benjamini & Hochberg procedure when appropriate, for example for all tests of item fit for a scale (Benjamini & Hochberg, 1995).

The Lavaan package (version 0.6-4) in the R statistical programming language (version 3.5.1) was used for the CFA analyses. Furthermore, the Diagonally Weighted Least Square estimation (DWLS; Li, 2016) was used since the items were ordinal. To assess the fit of the model to the data we used the Comparative Fit Index (CFI) and the Tucker Lewis index (TLI, also called the Non-Normed Fit Index, NNFI), with a value above .95 indicative of a fit (Hu & Bentler, 1999). The Root Mean Square Error of Approximation (RMSEA) and standardized Root

Mean Square Residual (SRMR) were also used to assess fit, with values below .06 and .08 indicating a fit, respectively (Hu & Bentler, 1999).

In addition to examining internal validity using IRT and CFA, we investigated external validity by calculating the correlation of the scale scores from the cognitive load measure with the single item cognitive load measure by Paas (1992), as well as learning outcome measures in each of the three studies. Furthermore, we conducted a stepwise regression with the learning outcome measure as the dependent variable in Study 2 and Study 3, to investigate if the proposed multidimensional conceptualization of extraneous cognitive load measure explained significantly more variance than the single item measure by Paas (1992).

Study 1: Investigating the Internal and External Validity of an Adapted Version of the Cognitive Load Scale in a Laboratory Experiment Sample

The overall objective of Study 1 is to evaluate the validity of a version of the CLS that was adapted to assess CL in a virtual learning environment. The construct validity of the scale is assessed through IRT and CFA and external validity is assessed through correlations with the single item cognitive load measure (Paas, 1992) and a measure of retention following a virtual learning session.

Sample Study 1

In Study 1, the sample consisted of 73 university students of which 44 were females and 29 were males. The students came from different fields of education. Most were psychology students (n = 34), however, many different fields were represented including anthropology, economics, dentistry, and medicine. The mean age of the sample was 23.97 with a standard deviation of 3.78. A total of 41 students (56.2%) reported that they had never used IVR before, 29 (39.7%) reported having tried it only a couple of times, 2 (2.7%) reported using it sometimes, and 1 (1.4%) student reported using it often.

Procedure and Materials Study 1

The experiment took place at an experimental lab at a large European university, and the participants were rewarded with a gift certificate equivalent of approximately 15 euros. The data was part of a larger study that investigated the efficiency of using IVR for learning. The learning material was embedded in an IVR simulation, which we developed in Unity3D, and optimized for the HTC Vive VR system. The simulation depicted a virtual hospital environment with several props including a doctor and a patient (see Figure 1). The learning material included 25

snippets of information about sarcoma cancer. The element interactivity was low, i.e. point and click interaction. The learning material included detailed information about sarcoma cancer taken from a national cancer society pamphlet. Participants were asked to learn as much as possible about sarcoma cancer in order to retell this information to a friend. After the VR experience, the participants filled out the adapted CLS (see Table 1) and conducted a retention test to assess how much of the knowledge they retained from the IVR learning experience. The measures of cognitive load were adapted to the context of learning about sarcoma cancer in immersive VR from Leppink et al., (2013). The retention test consisted of 25 multiple-choice questions, with one question representing each snippet of information about sarcoma cancer, which the participants were asked to remember.

Results for Study 1

Rasch Analyses Results for Study 1

The Rasch analyses of the three dimensions of cognitive load indicated no evidence against overall fit (see Table 2). This was the case for the tests of homogeneity, and for the individual item fits. An item-wise test of DIF in relation to the exogenous variables showed no DIF for gender, or age in relation to any of the items. Finally, the tests for local dependence were insignificant. Thus, overall we found that that each dimension fit the requirements of a Partial Credit Model and consequently the sum scores are statistically sufficient, and for all groups of the exogenous variables the scores are specifically objective.

Confirmatory Factor Analysis Results for Study 1

The CFA of a three-factor model including the three dimensions of intrinsic, extraneous, and germane load as separate factors achieved good fit. The CFI was 1.00 and the TLI was 1.00 which are above the required cut-offs of 0.95. The RMSEA was < .001 and the SRMR was .057, which are lower than the conventional cut-off of .06 and .08, respectively. In conclusion, the results provide support for the validity of the adapted version of the CLS by Leppink et al., (2013) in Study 1.

External Validity Results for Study 1

Table 4 provides an overview of the correlations between each of the dimensions from the adapted CLS (Leppink et al., 2013), the single item measure from Paas (1992), and the retention test score (See Table 3). The results indicate that the intrinsic CL scale had a significant negative correlation with retention (r = -.38), and the germane CL scale had a significant positive

correlation with retention (r = .36), but the correlation between the extraneous CL scale and retention (r = -.16) was not significant. Furthermore the correlation between Pass' item and retention was not significant (r = .05). Finally, although the directions of the correlations between the dimensions and the single item by Paas (1992) were as expected (positive correlations for intrinsic CL, and extraneous CL; and negative correlation for germane CL), none of these were statistically significant.

Discussion Study 1

Study 1 provided evidence for the internal validity of the adapted version of Leppink's (2013) CLS for virtual learning environments. All three scales fit the Rasch model, and CFA fit indices were good. Furthermore, the reliability coefficients for each dimension were high. Finally, the intrinsic cognitive load dimension had a significant negative correlation with retention, and the germane CLS had a significant positive correlation with retention as expected. The extraneous CLS had a negative correlation with retention but it was not significant. In conclusion, Study 1 provides initial evidence for the internal and external validity of the CLS for a virtual learning setting. Nonetheless, there were several limitations in this study, which call for further studies. The first major limitation in this study was a small sample size. The second limitation in the study was that it was a lab experiment, meaning that ecological validity could be lower than in a classroom setting. However, a strength of using a lab experiment is that it provides a high level of internal validity, which can be an advantage in terms of scale validation. Finally, the cognitive load measure developed by Leppink and colleagues (2013) only measured one component of extraneous cognitive load, which could limit its value for virtual learning environments.

Study 2: Developing and Investigating the Internal and External Validity of the Multidimensional Cognitive Load Scale for Virtual Environments (MCLSVE) in a University Sample

Study 2 was intended to build on the results of Study 1 by accounting for the limitations in that study. First, a thorough test development process was carried out to develop new items and scales for the expanded version of the Multidimensional Cognitive Load Scale (MCLS). Then, the internal and external validity of the new items and scales was tested in a sample of 140 university biology students who used an immersive VR simulation as part of their core education.

Development of the Multidimensional Cognitive Load Scale for Virtual Environments (MCLSVE)

To accommodate the nature of EL in virtual environments, the unidimensional extraneous CL scale was expanded into three separate scales pertaining to sub-dimensions of EL. These sub-dimensions have been highlighted by previous VR research as potential causes of extraneous CL, which could hinder learning (Makransky et al., 2019b). The original operationalization of extraneous cognitive load related to instructions and explanations from Leppink et al., (2013; 2014) was retained as the first sub-scale. The two new sub-scales of extraneous cognitive load included extraneous cognitive load related to interacting in a virtual learning environment, and extraneous cognitive load related to learning in a complex 3D virtual environment. These new sub-scales were proposed based on a number of studies, which have identified potential sources of CL when learning in VR environments (e.g., Makransky et al., 2019b; Meyer et al., 2019). A team consisting of a subject matter expert in educational psychology, a specialist in human computer interaction and IVR researcher, a graduate student in psychology, and a psychometrician developed new items. Finally, a team of four subject matter experts reviewed and revised the items (for a list of all initial items see Table 4).

Sample and Procedure Study 2

The participants were 140 first year biology students at a large European university who took an IVR simulation as part of their core education in the first semester of their education. A total of 57 were males, 80 were females, and three did not respond. The age ranged from 19 to 38 years of age, but most (> 80%) of the were students were between 21 and 26 years of age. A total of 83 (58.5%) students reported that they had not used IVR before, 46 (32.4%) reported that they had used IVR but for less than two hours, and 11 (7.7%) reported having used IVR for more than two hours. The IVR simulation consisted of several exercises regarding lab safety, for example: cleaning up a workplace and dealing with laboratory accidents such as an acid spill or getting acid in the eyes, identifying and learning key hazard symbols, and learning about other lab safety issues such as proper lab dress code (see Figure 2). There was a high element of interactivity, as students had to interact with different features of the virtual laboratory environment. They did this through a simple hand held remote. The difficulty of the tasks was low to medium. The simulation lasted for approximately 15 minutes and was developed by the educational technology company Labster. The lab safety simulation is further described by

Makransky and colleagues (2019a) and can be viewed on Labster's homepage (2020). The IVR simulation was administered on 30 Lenovo Mirage Solo headsets, with a small remote controller and headphones. Students had been scheduled to use the simulation during their lessons in groups of up to 30 students at a time.

Prior to using the IVR simulation students were asked to fill out a short questionnaire which contained several demographic questions. Following the lesson, students were asked to respond to a post-test questionnaire, which included the MCLSVE (see Table 4) and a retention test consisting of 15 items that measured declarative and procedural knowledge (e.g., "What should you always wear in the lab?"). The distinction into declarative and procedural knowledge was necessary because the learning activity in the IVR simulation included learning facts about laboratory safety such as safety symbols, as well as procedures such as the correct procedure for reacting when there is an acid spill.

Statistical Analyses Study 2

The statistical methods for Study 2 were similar to those used in Study 1 related to assessing fit to the PCM. Once fit to the model was established, a latent logit score was used to calculate the correlation with the single item cognitive load measure by Paas (1992) and the retention test score. Furthermore, a stepwise regression was used to investigate if the multidimensional conceptualization of extraneous cognitive load explained significantly more variance in the retention test than the single item measure.

Results for Study 2

Rasch Analyses Results for Study 2

The Rasch analyses of the five sub-scales of cognitive load from the MCLSVE indicated that three of the subscales exhibited evidence against fit to a pure PCM, the remaining two scales did not show any evidence against the fit of the PCM (see Table 5).

The intrinsic CL scale showed evidence of DIF relative to gender in the overall test, but not in the item-wise DIF test. However, when we tested for local dependence each time a significant instance of local dependence (LD) was added to the model a new one appeared until the model would no longer converge. Therefore, we conclude that this sub-scale did not fit the model, and we cannot assume that the IL scale in this sample is valid. We found no evidence of DIF relative or against homogeneity for the extraneous instruction load in the overall tests, and we did not find evidence of DIF during the subsequent item-wise DIF analyses. However, one

case of local dependence was found (between item 1 and 2 of the extraneous instruction load scale) and added to the model. Furthermore, item 2 ("The simulation really enhanced my knowledge and understanding of lab safety") was eliminated from the germane CL scale as it did not fit the model. A model without this item fit the PCM. For the remaining two scales we found no evidence against homogeneity, item fit, or local independence, or evidence for any form of DIF. Thus, the extraneous interaction, and extraneous environment cognitive load scales as well as the GL scale without item 2 fit the pure PCM. The reliability of fitting scales were satisfactory, ranging from .81 to .85.

Confirmatory Factor Analysis Results for Study 2

The fit of the data to a CFA with a second-order factor structure where cognitive load is measured by intrinsic, extraneous, and germane load, and extraneous load is further measured by instructions/explanations, interactions, and environment was investigated. The CFA indicated that the fit was greatly improved after eliminating item 2 from the germane CL scale. After eliminating this item, the results showed good fit to the model. Both CFI and TLI were 1.00, above conventional cut-off. RMSA was .013, and SRMR was .065, which are considered acceptable,.

External Validity Results for Study 2

Table 6 shows the correlations between the five sub-scales from the MCLSVE, the single item measure from Paas (1992) and the retention test. The table illustrates that the three extraneous CL sub-scales were significantly negatively correlated with retention (EL ins. = -.19, EL int. = -0.24, EL env. = -0.17); however, the germane CL scale and the single item cognitive load item from Paas were not significantly correlated to retention. Furthermore, the single item measure was not significantly correlated with any of the sub-scales from the MCLSVE.

A stepwise multiple regression was conducted to investigate the incremental validity of using the MCLSVE in addition to the single item measure by Paas in predicting retention scores. The regression revealed that at step one, the single item by Paas did not contribute significantly to the regression model ($F_{(1,136)}$ = 1,09, p = .298) and accounted for 0.8% of the variation in retention. Introducing the three adapted EL–scales showed significant improvement of the model ($F_{(4,133)}$ = 2,48, p = .047) and the model now explained 4.1% of variation in retention.

Discussion Study 2

All the scales, besides the IL scale, fit either a PCM or a GLLRM (which is a PCM with added local dependence or differential item functioning). The reliability of all the scales were satisfactory ranging from .81 to .85. Furthermore, significant negative correlations ranging from -.17 to -.24 were found between all of the EL sub-scales from the MCLSVE and retention. Finally, information from the latent scales provided significantly improved prediction of learning in a regression model and accounted for 4.1% of the variance in retention, compared to a model which only included the single CL item and only accounted for 0.8% of the variance in retention. Therefore, we found initial evidence for the multidimensional conceptualization of extraneous cognitive load within a specific virtual learning environment. Furthermore, we identified one item that did not fit the PCM for the GL scale but obtained fit to the model with the remaining three items. However, we failed to replicate the positive fit of the IL scale from Study 1. Although Study 2 provides initial validity evidence for the MCLSVE, one limitation was that the content of the simulation was easy, meaning that students may not have experienced a high level of cognitive load. Furthermore, most students did not have very much experience with IVR meaning that the novelty of the technology could influence the results. A further study was therefore necessary for assessing the generalizability of using the MCLSVE with a more difficult simulation in a sample of more experienced IVR users.

Study 3: Investigating the Generalizability of the Validity of the Multidimensional Cognitive Load Scale for Virtual Environments (MCLSVE) in a More Difficult Virtual Lesson

Study 2 provided some evidence of the internal and external validity of the MCLSVE within an IVR lesson that was relatively easy for the university sample as it was on the topic of laboratory safety, which was familiar to many students. Furthermore, the simulation only lasted roughly 15 minutes and the general level of reported intrinsic and extraneous cognitive load was relatively low (see means in table 7; IL = 2.1; EL ins = 1.61; EL int = 1.71; EL env = 1.65; GL =3.87). Therefore, Study 3 was designed to investigate the validity of the MCLSVE with the same participant pool in a more challenging learning environment. Thus, we investigated the validity of the MCLSVE with the same population of students but in a more challenging IVR simulation on the topic of gene expression, which was an integrated part of their biology education.

Sample and Procedure Study 3

The participants were 121 biology students from the same cohort as in Study 2. There were 55 males, 69 females, and 2 non-binary. All of the students had used IVR through their participation in Study 2. Furthermore, the learning material was novel and very difficult. Study 3 took place roughly one month after Study 2, and the learning intervention consisted of an IVR simulation "Gene Expression Unit: Use sequencing to unveil a gene linked to obesity" which was also developed by the educational technology company Labster (see Labster, 2020). The objective of the simulation was to learn how to sequence DNA using pig samples as a model, and to find out which genes in the DNA are linked to obesity. The simulation depicted a biology lab with equipment that learners could manipulate in order to complete the assignments (see Figure 3). The main tasks in the simulation were: preparing samples for Next-Generation Sequencing (NGS), understanding the principles behind the NGS technique and performing a qPCR experiment. The equipment and the questions for the cognitive load scales were identical to Study 2, however the retention test consisted of 10 multiple-choice items designed to assess the students' conceptual and procedural knowledge (e.g., "Why do we need to reverse transcribe the mRNA to cDNA before PCR and Sequencing?"). The analyses were also the same as in Study 2, however as an additional step, we analyzed the combined data from Study 2 and Study 3 to assess the construct validity of the MCLSV in a CFA.

Study 3 Results

Rasch Analyses Results for Study 3

Table 8 shows the results for the Rasch analyses of the five subscales of the MCLSVE, which did not identify evidence against overall fit. We did not find any violations of fit for the tests of homogeneity, or in the individual item statistics. Furthermore, there was no DIF for gender or age across any of the items. The reliability coefficients were also acceptable for all of the sub-scales with the exception of EL environment, which had a reliability of 0.68. However, there was an instance of local dependence in the EL interaction scale and one in the EL environment scale. These instances of local dependence were added to the model and the scales subsequently fit a GLLRM. Finally, item 2 was removed in the GL sub-scale as in Study 2 as it did not fit the PCM. Therefore, the results provide additional evidence of the validity of the MCLSVE when accounting for two instances of local dependence, and after one item that did not fit, was eliminated.

Confirmatory Factor Analysis Results for Study 3

The fit of the data from Study 3 was investigated using a CFA of the same second-order factor structure as in Study 2 after omitting item 2 from the GL scale. The CFI was 0.994 and TLI was 0.993. These are higher than the typical cut-offs of 0.95 which are considered good. Furthermore, the RMSEA was .063, and the SRMR was .072, which were close to the acceptable values of .06 and .08, respectively.

External Validity Results for Study 3

Table 9 provides an overview of the correlations between the five sub-scales in the MCLSVE, the single item CL measure and the retention test. Unlike in Study 2, none on the scales were significantly correlated with retention. Furthermore, a stepwise multiple regression with retention as the dependent variable revealed that at step one, the single item CL measure did not contribute significantly to the regression model ($F_{(1,117)} = 0.442$, $p = .507$) and accounted for 0.04% of the variation in retention. Introducing the three adapted EL–scales did not lead to a significant improvement of the model ($F_{(4,114)} = 0{,}553$, $p = .698$) and the model now explained 1.9% of variation in retention.

Discussion of Study 3

Study 3 provides further evidence of the construct validity of the MCLSVE. Most scales fit the Rasch model and the overall fit indices of the CFA were good. Only the EL interaction scale and the EL environment scale showed evidence of LD and none of the scales showed evidence of DIF. The reliability was estimated to .68 for the EL environment scale, which is below conventional cut-offs, however the reliability of all the remaining scales were acceptable.

None of the scales predicted retention. The reason that the EL sub-scales did not predict retention in Study 3 is unclear; however, one explanation might be a halo effect. Since Study 3 was more difficult than Study 2, the high level of intrinsic load experienced in the difficult task might also influence the reported extraneous load. A limitation in Study 3 as well as the other studies is that no objective measures of cognitive load were available to cross-validate the results.

Discussion of Differences across an Easy (Study 2) and Difficult (Study 3) IVR Lessons.

In a final analysis, we investigated if students reported cognitive load in an easy lesson (Study 2) was different from a more difficult lesson (Study 3), and if there would be differences across the extraneous cognitive load subscales in the MCLSVE. Table 7 reports the mean scores

on the MCLSVE across Study 2 and Study 3. The table illustrates that the independent samples t-tests indicated that students experienced significantly higher IL in Study 3 (M = 3.76, SD = 0.70) compared to Study 2 (M = 2.10, SD = 0.90) $t_{(259)}$ = 16.49, p < .001, d = 2.05. The same was true of the EL instruction sub-scale where students reported significantly higher mean CL in Study 3 (M = 2.39, SD = 0.75) compared to Study 2 (M = 1.61, SD = 0.61) $t_{(259)}$ = 9.23 , p < .001, d = 1.15. Students also reported higher EL interaction CL in Study 3 (M = 2.26, SD = 0.78) compared to Study 2 (M = 1.71, SD = 0.70) $t_{(259)}$ = 5.97, p < .001, d = 0.74, and EL environment load in Study 3 (M = 2.28, SD = 0.80) compared to Study 2 (M = 1.65, SD = 0.64) $t_{(259)}$ = 7.06, p < .001, d = 0.88. Likewise, the students reported significantly less GL in study 3 (M = 3.17, SD = .89) compared to Study 2 (M = 3.87, SD = 0.69) $t_{(259)}$ = 7.18, p < .001, d = 0.89. Therefore, these results indicate that the simulation in Study 3 resulted in higher perceived cognitive load across all sub-scales compared to the simulation in Study 2 and less germane resource for the simulation in Study 3 compared to the simulation in Study 2. One method for validating a test is to look for theory consistent group differences. Finding a higher level of intrinsic cognitive load for the difficult lesson (Study 3) compared to the easier lesson (Study 2) is consistent with the CLT and thus supports the validity of the scales.

General Discussion

The availability of virtual learning interventions is quickly increasing in pace with better accessibility to modern technologies such as virtual and augmented reality. Although virtual learning environments are highly attractive because they facilitate learning experiences that are otherwise too expensive, dangerous, or impossible in the real world (Bailenson, 2018), the instructional design decisions that need to be made when designing these environments are typically more complicated and underexplored compared to traditional media. Cognitive load has been repeatedly highlighted as an important factor for understanding learning in virtual learning environments (e.g., Makransky et al., 2019b; Meyer et al., 2019; Moreno & Mayer, 2002; Richards & Taylor, 2015). Accurately measuring cognitive load in relation to learning situations is essential for understanding when cognitive overload occurs, and for developing instructional design guidelines that can limit unnecessary cognitive load and facilitate learning. Although the current measures are useful for obtaining a general measure of students perceptions of CL, previous research has highlighted several limitations (e.g., Antonenko & Niederhauser, 2010; Kirschner et al., 2011; Makransky, et al., 2019c), and specifically called for the need to measure

different components of CL that are relevant when learning through, and designing virtual learning environments (e.g., Makransky, et al., 2019b).

In this paper, we conducted three studies. In the first study, we investigated the validity and reliability of an adapted version of the CLS (Leppink, 2013) for virtual learning environments within the frameworks of CFA and IRT. Overall, the results provide evidence of the validity of the measure in terms of construct validity through IRT and CFA, and convergent validity through correlation analyses. More specifically, in a lab setting where students learned about sarcoma cancer in IVR, we found that the three scales of IL, EL, and GL fit the PCM within the framework of IRT, and the CFA indicated that the data had excellent fit to the three-factor model.

In the second and third studies we further developed and validated the MCLS for Virtual Environments (MCLSVE) to include three sub-scales of EL that have been identified as important in virtual learning environments including EL from instructions, interactions, and the environment. The data fit the PCM or a GLLRM for the three EL sub-scales in both studies. In Study 2, the EL scales also had significantly higher predictive validity for retention, supporting the incremental validity of using the MCLSVE compared to the single item from Paas (1992). However, these results did not replicate in Study 3 where neither of the measures predicted retention.

The results from Study 2 and Study 3 suggested that item 2 (“The simulation really enhanced my knowledge and understanding of lab safety”) did not fit the GL scale. Therefore, future research should further investigate the validity of the scale by revising or eliminating this item. Furthermore, the IL scale did not fit the PCM in Study 2; however, the scale had perfect fit in Study 3. Therefore, future research should further investigate the validity of this sub-scale in different settings with other samples.

The CFA from Study 2 and 3 provided initial evidence for the construct validity of the MCLSVE. This is a major empirical contribution of this paper as it supports the multidimensional conceptualization of EL in virtual learning environments. Specifically, it suggests that the MCLSVE can be used to measure three separate components of EL consisting of EL from instructions, EL from interactions, and EL from the environment. Future research should further investigate the validity of the MCLSVE in other settings, with alternative learning

material and in different samples. Future research should also investigate the potential of other forms of EL that were not included in the MCLSVE.

Comparing the mean scores between Study 2 and Study 3 revealed meaningful group differences such that in Study 2 students reported lower IL, EL instruction, EL interaction, EL environment, and higher GL compared to Study 3, since the learning material in Study 2 was easier than in Study 3. Because IL and EL are proposed to be additive (Sweller et al., 2011), it follows that students would report higher levels of CL in settings where they have fewer resources available due to the difficulty of the material (IL). However, the results also highlight a potential limitation, namely that, even though the methods of interaction remained constant across study 2 and study 3. The students reported the interaction as more strenuous in study 3. Thus, the student's overall perception of high CL could also influence the EL interaction subscale, which theoretically should remain the same. While this is a limitation with the current scale it is a general limitation when using self-report measures (Makransky, Terkildsen & Mayer, 2019c). Future studies should therefore investigate the use of objective measures of CL and combine these with subjective questionnaires.

Implications

The results of this study, specifically Study 2 and Study 3, suggests that the use of a multidimensional framework for measuring ECL in relation to IVR is not only possible, but maybe also necessary to capture the complexity of extraneous load, which can exceed what is measured by a single scale or item. The new measures of ECL need to be investigated in a wider range of virtual simulations, which are outside the scope of this paper. However, further positive findings with the newly developed scales might lead to a further development of the theory of cognitive load in relation to virtual environments and learning in general. Future studies might also seek to implement multidimensional measures of the ECL outside of VRE by adapting the scales to assess different components of EL in traditional teaching environments.

Our findings suggest that the previously used single item scales provide very limited information compared to newer scales with multiple items. Scales with multiple items will typically provide higher fidelity in terms of having a larger range of sums scores. Furthermore, scales using multiple items can also be analyzed using IRT, as conducted in this article as well as others, which provide evidence for or against DIF across relevant groups, which is simply not

possible for single item scales. Finally, and specifically for the Rasch model family, we can assume that sum score is sufficient for the scales, which fit the pure Partial Credit Model.

In general, this paper provides a psychometric investigation of the validity of an adapted version of the CLS (Leppink et al., 2013) for virtual learning environments, and proposes the MCLSVE as an extension to the CLS that measures two additional sub-scales of EL. The three studies provide initial support of the validity of the adapted version of the CLS in a laboratory experimental setting (Study 1), and initial evidence of the validity of the MCLSVE in a high stakes higher education context (Study 2 and Study 3).

Open Practices Statement. The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

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

Antonenko, P. D., & Niederhauser, D. S. (2010). The influence of leads on cognitive load and learning in a hypertext environment. *Computers in Human Behavior*, *26*(2), 140–150. https://doi.org/10.1016/j.chb.2009.10.014

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.).

Baddeley, A. D. (1990). *Human Memory: Theory and Practice*. Allyn & Bacon.

Bailenson, J. (2018). *Experience on demand: What virtual reality Is, how it works, and what it can do* (1 edition). W. W. Norton & Company.

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.

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

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

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Comput. Hum. Behav.*, *25*(2), 315–324. https://doi.org/10.1016/j.chb.2008.12.020

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

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

Hambleton, R. K. (1991). *Fundamentals of item response theory* (1 edition). SAGE Publications, Inc.

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

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

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

Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. Cogent Education, 4(1), 1280256. https://doi.org/10.1080/2331186X.2017.1280256

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

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.

Labster. (2020). *Gene Expression Unit: Use sequencing to unveil a gene linked to obesity Virtual Lab*. Retrieved January 3rd 2020 from: https://www.labster.com/simulations/gene-expression-unit-2/

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

Leppink, J., Paas, F., van Gog, T., van der Vleuten, C. P. M., & van Merriënboer, J. J. G. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction*, *30*, 32–42. https://doi.org/10.1016/j.learninstruc.2013.12.001

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., Borre-Gude, S., & Mayer, R. E. (2019a). Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments. *Journal of Computer Assisted Learning*, *35*(6), 691–707. https://doi.org/10.1111/jcal.12375

Makransky, G., Lilleholt, L., & Aaby, A. (2017). Development and validation of the Multimodal Presence Scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior*, *72*, 276–285. https://doi.org/10.1016/j.chb.2017.02.066

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019b). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, *60*, 225–236. https://doi.org/10.1016/j.learninstruc.2017.12.007

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019c). 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

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

Meyer, O. A., Omdahl, M. K., & Makransky, G. (2019). Investigating the effect of pre-training when learning through immersive virtual reality and video: A media and methods experiment. *Computers & Education*, *140*, 103603. https://doi.org/10.1016/j.compedu.2019.103603

Moreno, R., & Mayer, R. E. (2002). Learning science in virtual reality multimedia environments: Role of methods and media. *Journal of Educational Psychology*, *94*(3), 598–610. Scopus. https://doi.org/10.1037/0022-0663.94.3.598

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

Parong, J., & Mayer, R. E. (2018). Learning science in immersive virtual reality. *Journal of Educational Psychology*, *110*(6), 785–797. https://doi.org/10.1037/edu0000241

Rasch, G. (1961). *On general laws and the meaning of measurement in psychology*. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Medicine. https://projecteuclid.org/euclid.bsmsp/1200512895

Richards, D., & Taylor, M. (2015). A comparison of learning gains when using a 2D simulation tool versus a 3D virtual world: An experiment to find the right representation involving the Marginal Value Theorem. *Computers & Education*, *86*, 157–171. https://doi.org/10.1016/j.compedu.2015.03.009

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

Sweller, J., Ayres, P., & Kalyuga, S. (2011a). 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. (2011b). *Cognitive Load Theory*. Springer Science & Business Media.

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

Table 1

Adaption of items text for Study 1

Scale	Leppink et al., 2013	Adapted version used in Study 1
IL	The topic/topics covered in the activity was/were very complex.	The topic covered in the simulation was very complex.
IL	The activity covered formulas that I perceived as very complex.	I perceived the treatment and diagnoses descriptions presented in the simulation as very complex.
IL	The activity covered concepts and definitions that I perceived as very complex.	The simulation covered concepts and definitions that I perceived as very complex.
EL	The instructions and/or explanations during the activity were very unclear.	The information display format used in the simulation made the content seem very unclear.
EL	The instructions and/or explanations were, in terms of learning, very ineffective.	The information display format used in the simulation made learning very ineffective.
EL	The instructions and/or explanations were full of unclear language.	The information display format used made the content in the simulation difficult to understand.
GL	The activity really enhanced my understanding of the topic(s) covered.	The simulation really enhanced my understanding of the topics covered.
GL	The activity really enhanced my knowledge and understanding of statistics.	The simulation really enhanced my knowledge and understanding of sarcoma cancer.
GL	The activity really enhanced my understanding of the formulas covered.	The simulation really enhanced my understanding of treatment procedures and diagnostic criteria related to sarcoma cancer.
GL	The activity really enhanced my understanding of concepts and definitions.	The simulation really enhanced my understanding of concepts and definitions about sarcoma cancer.

IL = Intrinsic Load, EL = Extraneous Load, GL = Germane Load

Table 2

Rasch analyses of the scales in Study 1

Scale	Overall test	Item fit	DIF gender	DIF age	LD	*r*
IL	✓	✓	✓	✓	✓	.80
EL	✓	✓	✓	✓	✓	.86
GL	✓	✓	✓	✓	✓	.91

Note. IL = Intrinsic Load, EL = Extraneous Load, GL = Germane Load, DIF = Differential Item Functioning, LD = Local Dependence, r = reliability.

Table 3

Correlations of Cognitive Load scales and Retention in Study 1

	A	B	C	D	E
A. IL	—				
C. EL	.21	—			
C. GL	-.21	-.55**	—		
D. Single item CL measure	.19	.20	-.23	—	
E. Retention	-.38**	-.16	.36**	.05	—

Note. IL = Intrinsic Load, EL = Extraneous Load, GL = Germane Load, * $p < .05$, ** $p < .01$.

Table 4

New items and scales included in the MCLSVE

Scale	
IL	The topic covered in the simulation was very complex.
IL	The simulation covered procedures that I perceived as very complex.
IL	The simulation covered concepts and definitions that I perceived as very complex.
EL ins	The instructions and/or explanations used in the simulation were very unclear.
EL ins	The instructions and/or explanations used in the simulation were, in terms of learning, very ineffective.
El ins	The instructions and/or explanations used in the simulation were full of unclear content.
EL int	The interaction technique used in the simulation was very unclear.
EL int	The interaction technique used in the simulation was, in terms of learning, very ineffective.
EL int	The interaction technique used in the simulation made it harder to learn.
EL int	The interaction technique used in the simulation was difficult to master.
EL env	The elements in the virtual environment made the learning very unclear.
EL env	The virtual environment was, in terms of learning, very ineffective.
EL env	The virtual environment was full of irrelevant content.
El env	It was difficult to find the relevant learning information in the virtual environment.
GL	The simulation really enhanced my understanding of the topics covered.
GL	The simulation really enhanced my knowledge and understanding of lab safety.
GL	The simulation really enhanced my understanding of the procedures covered.
GL	The simulation really enhanced my understanding of concepts and definitions.

Note. IL = Intrinsic Load, EL ins = Extraneous Load instructions, EL int = Extraneous Load interactions, EL env = Extraneous Load environment, GL = Germane Load.

Table 5

Results for the Rasch analyses of the scales in Study 2

Scale	Overall test	Item fit	DIF gender	DIF age	LD	*r*
IL	%[1]	✓	✓	✓	%[2]	-
EL ins	✓	✓	✓	%[3]	%[4]	.81
EL int	✓	✓	✓	✓	✓	.81
EL env	✓	✓	✓	✓	✓	.83
GL*	✓	✓	✓	✓	✓	.85

Note. IL = Intrinsic Load, EL ins = Extraneous Load instructions, EL int = Extraneous Load interactions, EL env = Extraneous Load environment, GL = Germane Load, DIF = Differential Item Functioning, LD = Local Dependence, r = reliability, * For the Germane Load scale item GL _2 was omitted to achieve a working model, 1 = significant difference from model for gender, 2 = non-convergence, 3 = Item DIF for age relative to item 2, 4 = LD gamma of 0.02 between item 1 and item 2.

Table 6

Correlations of Cognitive Load Scales and Retention in Study 2

	A	B	C	D	E
A. EL ins	—				
B. EL int	.73**	—			
C EL env	.68**	.83**	—		
D. GL	-.29**	-.41**	-.37**	—	
E. Cognitive Load scale single item	.08	.01	-.00	.11	—
F. Retention	-.19*	-.24**	-.17*	.05	-.09

Note. EL ins = Extraneous Load instructions, EL int = Extraneous Load interactions, EL env = Extraneous Load environment, GL = Germane Load, * $p < .05$, ** $p < .01$.

Table 7

Comparing mean scores from Study 2 and Study 3

	Study 2		Study 3			
Scale	*M*	*SD*	*M*	*SD*	*d*	*p*
IL	2.10	0.90	3.76	0.70	2.05	< .001
EL ins	1.61	0.61	2.39	0.75	1.15	< .001
EL int	1.71	0.70	2.26	0.78	0.74	< .001
EL env	1.65	0.64	2.28	0.80	0.88	< .001
GL	3.87	0.69	3.17	0.89	0.89	< .001

Note. IL = Intrinsic Load, EL ins = Extraneous Load instructions, EL int = Extraneous Load interactions, EL env = Extraneous Load environment, GL = Germane Load.

Table 8

Rasch analyses of the scales in Study 3

Scale	Overall test	Item fit	DIF gender	DIF age	LD	*r*
IL	✓	✓	✓	✓	✓	.82
EL ins	✓	✓	✓	✓	✓	.78
EL int	✓	✓	✓	✓	%[1]	.81
EL env*	✓	✓	✓	✓	%[2]	.68
GL**	✓	✓	✓	✓	✓	.89

Note. IL = Intrinsic Load, EL ins = Extraneous Load instructions, EL int = Extraneous Load interactions, EL env = Extraneous Load environment, GL = Germane Load, *DIF* = Differential Item Functioning, LD = Local Dependence. * A GLLRM was achieved by omitting Item O (EL_env_1), ** For the Germane Load scale item GL_2 was omitted to achieve a working model, 1 = LD gamma of .01 between item 1 and item 2, 2 = LD gamma of .20 between item 2 and item 3.

Table 9

Correlations of Cognitive Load scales and Retention in Study 3

	A	B	C	D	E	F	G
A IL	—						
B. EL ins	.11	—					
C. EL int	.08	.63**	—				
D EL env	.11	.45**	.62**	—			
E. GL	-.24**	-.48**	-.60**	-.41**	—		
F Cognitive Load scale single item	.19*	-.14	-.26**	-.16	.24**	—	
G Retention	.02	-.13	-.09	-.05	.03	.09	—

Note. IL = Intrinsic Load, EL ins = Extraneous Load instructions, EL int = Extraneous Load interactions, EL env = Extraneous Load environment, GL = Germane Load, * $p < .05$, ** $p < .01$.

Figure 1 Screen shot from study 1 regarding Sarcoma cancer depicting a doctor and a hospitalized patient in a bed

Figure 2 Screen shot from study 2 regarding lab safety. The user's tasks is to identify potential safety hazards in the lab environment

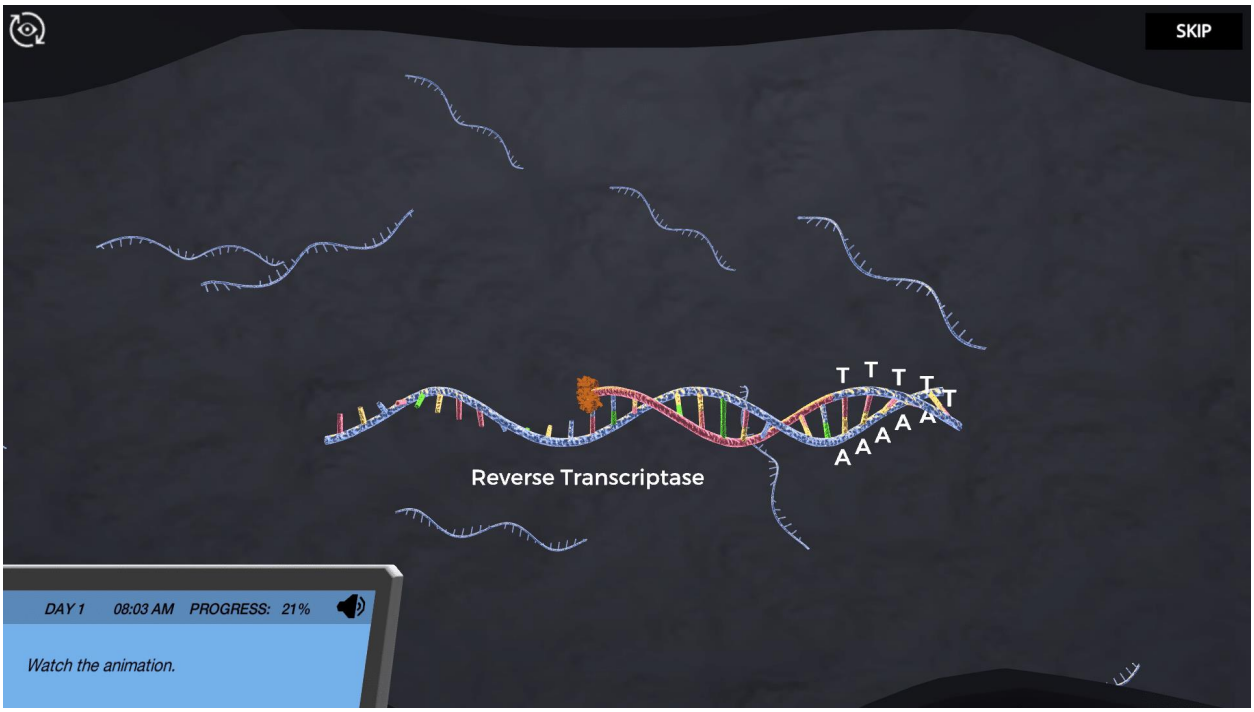

Figure 3 Screen shot from study 3 about gene expressions. The user is watching an animation about reverse transcriptase