TOWARDS A COGNITIVE THEORY OF MULTIMEDIA ASSESSMENT (CTMMA)

Kirschner, P.A.[1], Park, B.[2], Malone, S.[2], & Jarodzka, H.[3]

[1] *paul.kirschner@ou.nl Open University of the Netherlands & Oulu University, Finland*

[2] *b.park@mx.uni-saarland.de Saarland University, Germany*

[2] *s.malone@mx.uni-saarland.de Saarland University, Germany*

[3] *halszka.jarodzka@ou.nl Open University of the Netherlands & Lund University, Sweden*

Abstract

Much is known about assessment in all its forms and the corpus of theory and knowledge is growing daily. In a similar vein, the use of multimedia for learning also has a sound basis in research and theory, such as the Cognitive Load Theory (CLT; Sweller, Van Merriënboer, & Paas, 1998), human information processing (e.g., Atkinson & Shiffrin, 1968; Miller, 1956; Paivio, 1986), and praxis in the form of evidence-informed design principles often based on the Cognitive Theory of Multimedia Learning (CTMML; Mayer, 2005b). However, the combination of the two lacks both theoretical underpinnings and practical design principles. Multimedia assessment (MMA) is, at best, either a translation of paper-based assessment and assessment principles to the computer screen or an attempt to make use of the theory and principles underlying multimedia learning (i.e., CTMML). And this is the problem. In the first place, MMA needs, just as Multimedia Learning (MML), its own theory and principles. Just as MML was not simply the translation of paper-based learning to the computer screen, MMA requires its own place. In the second place, the application of CTMML and its principles to assessment leads to problems. The CTMML is based upon the idea that learning should be facilitated by the proper use of CTMML principles and its underlying theories (CLT, Human Information Processing). In cognitive load terms: germane load is increased while extraneous load is avoided so as to facilitate effective and efficient learning. But the goal of assessment is not learner facilitation, but rather separating the wheat from the chaff. Those who do not possess the knowledge and skills need to not be able to answer the question while those who do have the knowledge and skills need to answer correctly. This may mean that certain forms of extraneous load need to be increased while germane load needs to be minimised. This chapter will kick off the road to a Cognitive Theory of Multimedia Assessment (CTMMA).

Key words

Assessment; Multimedia; Instructional design; Cognitive load

Introduction

In education and training using both paper-based and computer-based learning materials we see both a convergence of opinions on and an adoption of instructional design principles and practices for their use. Instructivists and constructivists have found a certain degree of common ground in most if not all of the guidelines and principles found in the Cognitive Theory of Multimedia Learning (CTMML; Mayer, 2001) and Cognitive Load Theory (CLT; Sweller et al., 1998). These guidelines and principles, arising from paper-based instructional materials have been expanded and specified for the increasing use of computers and computer-based instructional materials and learning environments and specifically multimedia materials and learning environments. Multimedia is defined here as a combination of text, audio, still images, animation, and/or video content.

Now that multimedia learning materials have become commonplace and educators, trainers, instructional designers, and educational policy makers (including politicians) have embraced the ability of such materials to personalize teaching, training, and learning leading to more effective, efficient and possibly more enjoyable teaching and learning experiences, a concomitant increase in attempts is observable with respect to designing, developing, and implementing multimedia assessment (e.g., in the Netherlands: De Boer, 2009; in Germany: Hartig & Klieme, 2007; Dennick, Wilkinson, & Purcell, (2009), Hamm & Robertson, 2010). We have deliberately chosen for the word ‘assessment’ and not for ‘testing’ as assessment is, in the context of this chapter, a much broader concept which includes testing. Testing (sometimes called examination) is almost always used in a summative way to determine what someone knows or has learned. Testing is actually subsumed by assessment in that it is a form of assessment which is intended almost exclusively to measure a test-taker's knowledge, skill, aptitude, physical fitness; in other words it classifies a person assigning her/him a level or score. Assessment expands this to include the process of measurably documenting the progress of the learner (i.e., her/his knowledge, skills, attitudes and beliefs) in measureable terms to make improvements in and help guide that process. And here is where one can encounter problems that are counterintuitive, counterproductive, and possibly detrimental to assessment. While the design and use of multimedia for instruction is based upon sound and often tested theories (i.e., CTMML, CLT) with concomitant guidelines, using multimedia for computer-based assessment (CBA) is not.

On the one hand, CBA is often based upon traditional design principles that have been developed and tested – for better or for worse - for paper-based applications, which are quite limited as compared to CBA with regard to presentation and response formats. The question is whether tried and tested instructional guidelines can simply be transferred to multimedia assessment and which aspects of CBA require their own proper principles. On the other hand, some designers use the CTMML and/or the CLT for the design and development of assessment. Different indicators are used for measuring cognitive processing during learning and its measurable immediate or delayed consequences (for an overview see also Brünken, Seufert, & Paas, 2010; Van Mierlo, Jarodzka, Kirschner, and Kirschner (2012). These include subjectively self-rated cognitive load or mental effort (Leppink, Paas, van der Vleuten, van Gog, & van Merriënboer, 2013; Paas, 1992), objectively measured cognitive load via the dual-task paradigm (Brünken, Plass, & Leutner, 2004; Brünken, Steinbacher, Plass, & Leutner, 2002; DeLeeuw & Mayer, 2008; Park & Brünken, 2015), cognitive load measured with different eye-movement phenomena (Jarodzka, Jansen, Kirschner, & Erkens, 2015; Knörzer, Brünken, & Park, 2016; Marshall, 2002; Mayer, 2010; Park, Knörzer, Plass, & Brünken, 2015; Park, Korbach, & Brünken, 2015), and different levels of learning performance distinguishing, for example, between retention and transfer performance (e.g., Marcus, Cooper, & Sweller, 1996) or knowledge about processes and structures (Park, Münzer, Seufert & Brünken, 2016) or combined measures. The question here is whether principles meant to make learning from multimedia materials effective and efficient can be directly transferred to and used in an assessment situation since the goals of learning and assessment are different from and may even conflict with those of learning. For example while the goal of using certain learning materials or types of learning materials might be to reduce extraneous cognitive load so as to facilitate learning (Sweller et al., 1998), the goal of introducing assessment or types of assessment materials might be to increase extraneous cognitive load so as to better distinguish between novices and experts. In this regard, the International Test Commission refers in their guidelines on computer and internet testing to the use of advanced multimedia features in assessment, stating, that these "should be used only where justified by validity" (p. 147).

To overcome this research gap, the present article introduces and defines a cognitive theory of multimedia assessment (CTMMA) and presents the first derived principles for the design of multimedia assessment materials.

Birth and Growth of Computer-Based Assessment

With increasing technical development, the use of computers for assessment is inevitable to take over assessment in general. Though original computer-based assessment was actually only computer-based testing (i.e., was only designed and used for making summative decisions) we will still use the term assessment. Educational researchers and designers should be prepared to deal with this and be able to provide guidelines while CBA is being introduced instead of only reacting to students being confronted with bad design in CBA. Thus, the question is not about media whether assessment should be on paper or on computers, but instead about the methods: How to design CBA so that it does not hamper students' performance nor its assessment, but instead uses its full potential to capture students' level of knowledge, skills and potential performance as adequately as possible so as to facilitate their progress in learning, skills attainment, and attitude adoption. Avoiding incorrect diagnostic decisions is an important goal for assessment for many reasons. Inappropriate assessment can even affect safety concerns. Theoretical driving tests have, for example, been recently adapted to computer based assessment systems in many countries in the world. As this test is to assess aspects of driving competence, its function is to identify those applicants who are not yet competent enough to drive safely and therefore need further training. Besides, high norms for assessment are also required in less standardized contexts of assessment, for example, in multimedia learning studies. Appropriate assessment of learning outcomes is required for conclusions about the effects of different instructional methods, adaptive instruction and adaptive assessment, and real-time feedback.

To address this methodological question, the advantages and the dangers of CBA due to the media change from paper-based to computer-based and due to different possibilities that CBA affords are considered next. One crucial change coming from CBA is that it allows for automated assessment and analysis of outcomes. This automated assessment requires less time to prepare the tests for their administration (i.e., there is no need to prepare hundreds of print-outs), makes it easier to prepare parallel forms so that students sitting next to each other cannot just copy what their smarter neighbors have answered, and requires less time to evaluate their test performance. As ‘time is money’, CBA can have – even at this very basic level – a distinct and measurable economic advantage over its paper-based peer (c.f. Jurecka & Hartig, 2007). Furthermore, CBA ensures quality criterions by providing enhanced standardization of administration, scoring, and interpretation of the results. The automated analysis of the test performance leaves less room for careless mistakes made by teachers/markers and is thus more reliable and valid. In addition, CBA makes the analysis of the test as a whole (e.g., test-retest reliability, item analysis, item-test reliability, et cetera) quicker and easier. Moreover, CBA allows for adaptive assessment that is for example recommended for adaptive learning systems when considering an individual difference perspective concerning prior knowledge (i.e. Kalyuga, Ayres, Chandler, & Sweller, 2003) or learner characterisitics like spatial ability (i.e. Korbach, Brünken, & Park, 2016; Münzer, 2012, 2015; Münzer, Seufert, & Brünken, 2009; Park, Korbach, & Brünken, 2015; Park et al., 2016). With a large enough database of items that are well-designed accompanied by an adaptivity algorithm, it is possible to easily provide different versions of a test to different groups of learners in different situations. Hence, the assessment can be adapted to each student’s knowledge level and thus, not only be conducted more quickly (by avoiding too difficult and too simple questions for specific learners), but also be more accurate by carving out the abilities of a student in detail. Added to this, CBA allows the use of very different forms and combinations of media to students (e.g., sound, video, animation) that may represent a certain task better than only text and static pictures – often also only in black and white - such as in paper-based assessment (PBA). Besides the presentation of various stimuli, CBA offers the opportunity to record aspects of the participant's behavior (e.g., response times) that cannot be logged by the means of PBA. Finally, CBA can be used in different places at different times and thus reach students across the world without the need to be physically present at a certain place at a certain time.

However, the possibilities of CBA can be negated and/or even have serious disadvantages. For instance, when introducing CBA one can easily be tempted to simply put a PBA on a computer. This change of medium without adaptation to it can, for example, cause disadvantages to processing the information (e.g., paper pages that can easily be turned vs. computer pages that cannot be revisited) or to responding to task demands (e.g., using a pen vs. using the keyboard). Especially, for speed tests these difficulties that come along with CBA can lead to a test bias (i.e., participants, who rarely use a computer can be at a disadvantage). A second problem is that, since there are no explicit guidelines for CBA, the technical possibilities that it makes possible can easily lure test designers to implement advanced media (e.g., 3D visualizations, hypertext links), just because it is possible, without considering the consequences of this for its demands on processing and on test reliability and validity. This can lead to what designers call "Christmas tree" designs with lots of colorful trinkets and candy hanging on it taking away from the functionality needed. Finally, the ultimate goal of assessment is to reliably and validly distinguish someone who knows something or can do something from someone who cannot / does not and/or determine who is a novice, who is an expert, and where someone is on the continuum between the two. Improper use of the possibilities of multimedia in the assessment situation can easily lead to false positives (i.e., Type I errors where a learner who lacks the knowledge and skills is classified as having them because the multimedia made the test items answerable with the required prerequisite knowledge and/or skills) or false negatives (i.e., Type II errors here a learner who has the knowledge and skills is classified as not having them because the multimedia and their use made the test items unanswerable / unreliable).

Traditional assessment and item design

According to Lienert (1969); (or more recently: Moosbrugger & Kelava, 2012) a test is a scientific routine to examine one or several personality features to make a quantitative statement about the relative degree of this feature's characteristic. Such assessments can measure very different aspects. As the present paper focuses on *performance* (cf. Bortz & Döring, 2013), each assessment has to meet three quality standards, that is, objectivity (i.e., different coders must come to the same results), reliability (i.e., when repeating the assessment under similar circumstances similar outcomes must be reached), and validity (i.e., individuals with a similar degree of a feature characteristic must come to a similar outcome). The latter is considered to be very important, as it demands evidence for a strong relation between the construct that is proposed to be assessed and the features that are actually assessed. Among others, two aspects of validity can be distinguished: criterion and ecological validity. Criterion validity refers to the relation between score and a certain criterion beyond the assessment situation. Ecological validity concerns the extent to which the assessment demands are similar to the demands of typical tasks in the respective domain.

The process of constructing an assessment can be divided in six phases, namely planning the assessment, design of the assessment + construction of the assessment items, analysis of assessment items, exploitation of item analysis, empirical testing of assessment quality criteria, and standardization of the assessment (Lienert & Raatz, 1994). The present paper focuses in particular on the second phase, that is, the design of the assessment and the construction of the assessment items. An item consists of two parts: a stimulus part and a response part.

From principles on item design many guidelines can be drawn on how to formulate questions or statements given to the participant (e.g., not to use ambiguous terms), which response modes to choose when (e.g., multiple-choice questions vs. open answer formats), and how to compose several items into one test (e.g., according to discrimination power). However, though certain principles on the actual *layout design* of items exist for paper-based items, there are actually no principles for multimedia-based items. This is surprising as multimedia has the powerful potential to increase ecological validity of a test, because it can reflect and/or simulate many aspects of real-life tasks in more detail than traditional paper-based assessment (PBA). On the other hand, research, which is derived from or refers to cognitive theories on learning (see next section), shows that the use of multimedia instruction must be very carefully considered and the design of such multimedia material must take human cognitive architecture into account to not hamper performance.

Within traditional item development, there is a focus on advantages and disadvantages of different response formats. Tasks with closed response formats (e.g., multiple choice questions, matching tasks) are on the one hand easy to evaluate, ensure high objectivity, and are economic for different

reasons. On the other hand, such formats are often difficult to create (e.g., presenting the necessary number of plausible alternatives in a multiple choice test) and aspects of their validity are questionable (e.g., in a car driving exam the answers of the multiple choice questions can often be simply learnt by heart). Advantages of open response formats (e.g., open questions, essays) are high content validity and easy development. Disadvantages are uneconomic (e.g., evaluation takes a lot of time, the need for a second assessor) and unstandardized evaluation (e.g., that two or more assessors reach the same evaluation of the answer). Beyond these classical response modes, items may be more authentic, in terms of context authenticity or the authenticity of the response formats (Meyer, 1992). Those formats that are most ecologically valid, are often also most unstandardized. As all these kinds of response formats have pros and cons that are mutually exclusive in PBA, the decision for a certain response format always involves a conflict. This dilemma can be addressed by multimedia assessment, which allows for standardized and ecologically valid assessment at the same time.
Ecological validity can be achieved by providing authentic sensations (e.g., animations instead of still pictures to visualize motion or sound instead of pronunciation notation) and realistic tasks (e.g., simulations). At the same time, internal validity can be assured by standardized test implementation, task assignment, and interpretation of results, which would be only possible to a limited extent within authentic assessment. CBA should make use of this advantage to provide valid assessment at all points.

Limitations of Cognitive Theories on Multimedia Learning when Applied to Assessment

Even though not much is known on how to design computer-based multimedia assessment (CBMMA) environments, one possible source of recommendations and guidelines can be found in those principles that are available for the design of multimedia in computer based instruction, though this must be done with caution. Two leading theories in this field of research are the Cognitive Theory of Multimedia Learning (Mayer, 2001, 2005b) and the Cognitive Load Theory (Sweller et al., 1998). Both theories are based on similar assumptions that lead to similar recommendations for instructional design of learning material. What do these theories tell us about the idea of learning and why does this not work for assessment?

Cognitive Theory of Multimedia Learning (CTMML)

CTMML is targeted on creating a plausible theoretical construct, which is consistent with known principles of research on learning and instruction. CTMML (Mayer, 2001, 2005b, 2009) is based on three assumptions, as already summarized in numerous publications like, for example, by Park (2010) as follows. The first assumption, which is a main assumption for many cognitive theories comprises that human working memory, the cognitive subsystem for processing current information, is limited in its capacity for processing (Baddeley, 1992; Chandler & Sweller, 1991; Miller, 1956; Miyake & Shah, 1999). The second assumption is that meaningful learning requires active processing of information by the learner (cf. Figure 2). For active processing different cognitive processes are necessary such as focusing the attention on the relevant learning content (i.e., selection), mentally organizing information in a coherent way (i.e., organization) and integrating new information with existing knowledge (i.e., integration). These three essential cognitive processes result in the so-called SOI-model (Selection-Organization-Integration; Mayer, 1996) summarizing active processing of an engaged learner. The last assumption of the CTMML is the dual channel assumption (cf. Figure 1), which is derived from the Dual-Coding Theory of Paivio (1986). Two channels of information processing have to be differentiated: verbal information is processed in the verbal/auditory channel, while pictorial information is processed via the visual/pictorial channel and limited capacity is assumed for each channel. In detail, active processing of pictures and words begins with the perception of these external representations via sensory memory. After that, the selection of relevant information begins within the working memory and results by means of an organization process in pictorial or verbal mental models. These internal representations are integrated by an active integration process to a coherent mental model ending in storage in long-term memory.

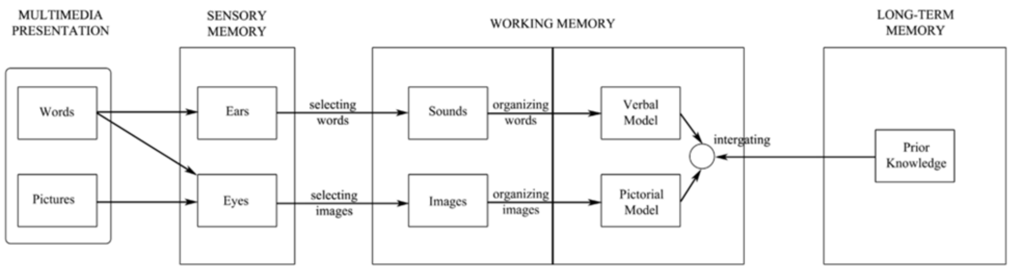

Figure 1. Cognitive Theory of Multimedia Learning (Mayer & Moreno based, in part, on the Dual-coding Theory of Paivio).

In sum, it is possible to empirically test hypotheses, which can be derived from CTMML. This is what Mayer and other researchers successfully showed, documented in three handbooks of multimedia learning (Mayer, 2001, 2005a, 2009). In the most recent version of his handbook of multimedia learning, Mayer (2009) distinguishes between principles for reducing extraneous cognitive processing (i.e., coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principle), principles for managing essential cognitive processing (i.e., segmenting, pre-training, and modality principle), and principles for fostering generative processing in multimedia learning (i.e., multimedia, personalization, voice, and image principle). These principles, however, cannot be simply translated to assessment. For instance, one essential principle that is derived from CTMML is about coherence of the learning material. As, according to CTMML, learning consists in developing a coherent mental representation of the learning contents, this can be fostered best by avoiding incoherence in presented learning materials. The corresponding reasonable deduction from CTMML for multimedia assessment is that tasks for assessments are to assess, whether the student has a correct coherent mental model. However, the question that arises in this context now is, whether the assessment materials should be designed to be coherent, too. With the present paper, it is hypothesized that the opposite assumption might be true. Criterion validity of a task is expected to be higher if incoherencies appear within the assessment materials. Dealing with incoherencies can be an indicator of competence; because of their coherent mental model experts can compensate or block out incoherencies. Similar questions arise when trying to transfer design principles that have been derived from CLT to assessment principles as described in the following section.

Cognitive Load Theory

Like many working memory models (Baddeley, 1992; Mayer, 2001; Paivio, 1986), CLT assumes that the capacity of working memory is limited and therefore learning is hampered when capacity is exceeded. In detail, CLT assumes that: (1) different learning issues can be distinguished by complexity of the learning task, (2) human working memory, the cognitive subsystem for processing current information, is limited in its capacity for processing (Baddeley, 1992; Miyake & Shah, 1999); and (3) learned content is stored in capacity unlimited long-term memory by using meaningful structured complex mental representations, in the form of schemata (Rumelhart & Ortony, 1976; Schank & Abelson, 1977).

The description of CLT exists already in numerous publications and the following one is out of a paper on cognitive and affective processes in multimedia learning by Park, Flowerday, and Brünken (2015). CLT (Kirschner, 2002; Plass, Moreno, & Brünken, 2010; Sweller, Ayres, & Kalyuga, 2011) assumes that knowledge acquisition depends on the efficiency of the use of available (limited) cognitive resources. The extent of cognitive load is thereafter determined by three components. First, *intrinsic cognitive load* (ICL) is related to the complexity of the learning content in terms of number of elements and the interactivity between those elements. Thus, intrinsic load depends on the number of elements and the relationships between them that must be simultaneously processed in working memory to learn the material being taught. The larger the number of elements of the material that needs to be learned and the higher the interactivity of those elements, the higher the intrinsic load of the material. Second, *extraneous cognitive load* (ECL) is caused by the cognitive demands imposed by instructional design that is not conducive to learning. The better the learning material is presented, considering the cognitive architecture and empirically proved instructional design principles, the lower the extraneous cognitive load. Instructional material, which does not specifically lead to learning and/or distracts from learning (e.g., search behavior which is not part of the learning goal) should thereafter be avoided. Finally, *germane cognitive load* (GCL) is the load that results from engaging in learning activities that effectively and efficiently foster schema acquisition. Germane cognitive load is thereafter also elicited by instructional material that facilitates or is beneficial for effective and efficient learning processes and therefore beneficial for the learning outcome. Whereas extraneous sources of load hinder learning, intrinsic sources of load reflect the complexity of the given learning task in relation to the learners´ level of expertise, and germane sources of load promote learning by helping students engage in the process of schema formation and automation. A basic assumption of CLT is that the total cognitive load experienced during learning is additively composed of these three load types, the so-called additivity hypothesis (Moreno & Park, 2010). If total cognitive load is excessive, learning and problem solving will be inhibited. The triarchic model of CLT is shown in Figure 2 that is adapted from a summary on the historical development of CLT by Moreno and Park (2010).

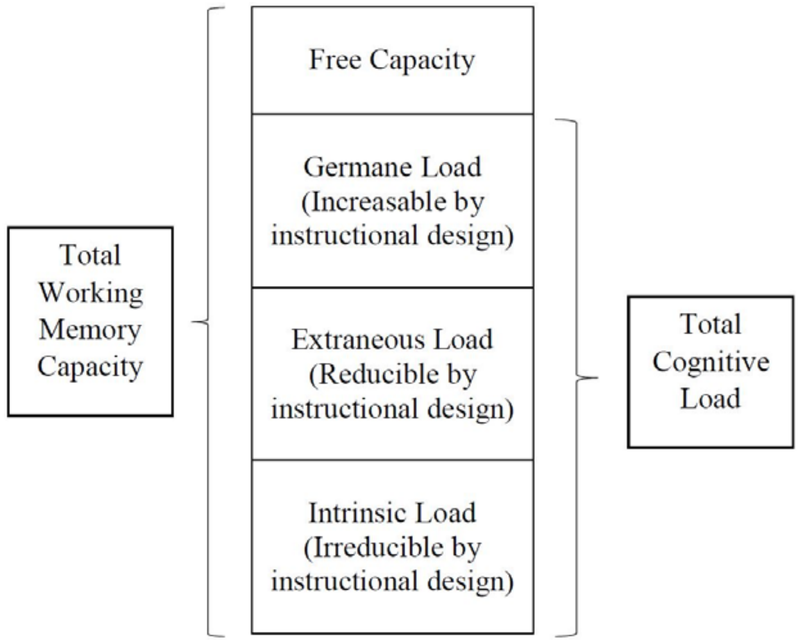

Figure 2. Triarchic model of cognitive load theory (adapted from Moreno & Park, 2010, © Cambridge University Press, reprinted with permission).

In sum, CLT results in the practical implication that extraneous cognitive load can be reduced by optimization of instructional design in order to free-up capacity for germane cognitive load. Especially reducing extraneous load is therefore assumed to facilitate learning. For the present paper, the question arises if the principle to reduce extraneous load is also relevant for and transferable to CBMMA.

Is this the good approach to assessment?

Both theories, CTMML and CLT, assume that the human information processing system is limited in terms of capacity and durability. As a consequence, they recommend minimizing the amount of information that needs to be processed at any one time. A second joint assumption is that information of different modalities is initially processed in different parts of the human information processing system. Hence, to make optimal use of this limited system, both parts should be used. The third assumption is that for information to be stored durably, it must be processed actively. Optimal learning material should support active processing of the to-be-learned information. All these joint assumptions are necessary and relevant when investigating information processing during learning and instruction. However, when looking from the other side of the coin on information processing that is from the retrieval of stored information, other relevant aspects need to be considered. With the goal to create a plausible theoretical construct, which is consistent with known principles of research on CBA, all principles that have been derived from CTMML have to be proved carefully with respect to its suitability for CBMMA. As already mentioned, criterion validity of a task is for example expected to be higher if incoherencies appear within the assessment materials.

Moreover, the following question arises when considering CLT from the other – assessment - side of the coin: Is there such a thing as intrinsic, extraneous and germane assessment load and if so, what are they? And, is reducing extraneous cognitive load the right thing to do for assessment? A constructive dilemma exists between fostering instructional understanding by reducing extraneous load and ensuring ecological validity in assessment by keeping this load relatively high. Fostering instructional understanding can be achieved by reducing extraneous load, which is essential for learning as well as for assessment, as aspects of reliability and validity can be ensured because the measuring error is being reduced. However, for some tasks, especially within the assessment of complex skills, minimizing extraneous cognitive load would mean reducing the task's complexity and, at the same time, making it less similar to (i.e., more discriminable from) the tasks that usually exist in the specific domain (i.e., low ecological validity). This would make it highly problematic – if not impossible - to determine whether the assessee has acquired the knowledge, skills, and/or competencies required.

In sum, multimedia principles derived from CTMML need to be varied or even reversed in most cases of CBMMA. In addition, design principles, which are derived from CLT seem not to be simply transferable to CBMMA in the same way. For instance, the ways to reduce cognitive load in multimedia learning described by Mayer and Moreno (2003) as principles to foster learning can appropriately be used in assessment when varying cognitive load in these ways for two purposes: ensuring ecological validity in assessment by keeping Cognitive Assessment Load (CAL) relatively high and opening the possibility to test the limits of learners by varying CAL (from low until high) within the CBMMA. Thus, the following multimedia effects explicated by Mayer and Moreno (2003) could be used to vary CAL: modality, segmentation, pre-training, coherence, signaling, spatial contiguity, redundancy, temporal contiguity, voice, and personalization. In detail, the variation of CAL can be assigned to different kinds of assessment load:

1. Intrinsic Assessment Load (IAL) varies, for example, by using more or less complex assessment tasks or by using differing amounts of pre-training or explanations to already known labels or procedures of the assessment material that could differ from learning material as it is often the case in transfer assessment tasks;
2. Extraneous Assessment Load (EAL) associated with most of the principles mentioned varies, for example, by using incoherencies or redundant assessment material or additional load by using for example dual-task methods;

3. Germane Assessment Load (GAL) could vary, for instance, by using animating material to foster the learners' assessment performance such as positive feedback or other methods, which increase the learners' engagement within the assessment situation.

Design Principles in Multimedia Assessment

On the basis of CTMML and CLT various principles to guide instructional design have been formulated and empirically studied. Can design principles that originate from learning be appropriate for assessment?

Tasks usually consist of two parts: a stimulus part and a response part. Both parts of a task have different functions: Whereas the stimulus part relates to information presentation, the response part is about what kind of reaction is demanded from the participant. It should be discussed for the both parts of a task separately, whether the use of instructional design principles for the purpose of designing tasks for assessment is promising.

Applying cognitive principles to the presentation of test items

Regarding the stimulus part of a task, there is clear evidence against the adoption of instructional guidelines to assessment. The *expertise reversal effect* (Kalyuga et al., 2003) – where instructional techniques that are highly effective with novices lose their effectiveness and even have negative consequences when used with more experts and vice versa -can be interpreted as an indicator for the inappropriateness of many design principles for assessment. In particular, the criterion validity of a test is expected to be threatened by an uncritical adoption of the common multimedia learning design principles to multimedia assessment situations.

An empirical indicator for criterion validity is a clear performance difference (with lower total load) in favor of domain experts compared to novices. According to the expertise-reversal effect, the application of some design principles in multimedia learning support novices and hamper experts. This question is also related to the question on cognitive load posed on an assessee: How can an *optimal* level of extraneous load be reached that allows instructional understanding *and* ecological validity at the same time?

According to general expertise research, experts clearly outperform novices in a specific domain: they solve problems faster and make fewer mistakes (Ericsson, Charness, Feltovich, & Hoffman, 2006; Posner, 1988). Thus, for CBA this means that adding time-pressure (or logging time-on-task) to the environment may help to distinguish between individuals of different levels of expertise. In terms of CLT this expertise-related difference in performance is expected to be caused by different amounts of intrinsic and germane cognitive load in experts as compared to novices, which in turn is due to a difference in knowledge structuring. While intrinsic assessment load is the load that arises from the subjective difficulty of a certain task, germane load in assessment can be defined as load that is produced by the processes of information retrieval and problem solving.

Experts and novices differ with respect to intrinsic load: Experts have more prior knowledge than novices as well as having this prior knowledge organized differently in their schemata (i.e., they have larger and more complex schemata which function as one chunk or information element). For some domains, this knowledge structuring is not only encapsulating in continuously larger chunks, but this structure is entirely different than the one of novices or even intermediates (e.g., in medicine: Boshuizen & Schmidt, 1992). The same task, thus, is expected to be more difficult for novices than for individuals with higher expertise (e.g., intermediates, experts). A good task for skill assessment is expected to reveal this difference in intrinsic load: experts are assumed to be able to effectively and efficiently solve a complex task while novices are assumed to fail. Especially for less demanding tasks (low intrinsic load), an optimal level of induced extraneous load can support assessment. With an optimal amount of extraneous load experts still have free resources to accomplish the tasks while the novices' complete cognitive capacity will be consumed by intrinsic and extraneous load (cf. Figure 3).

Figure 3. Optimal division of intrinsic (IAL), extraneous (EAL), and germane assessment load (GAL) in tasks for multimedia assessment

Basically, real experts must be able to perform also under sub-optimal circumstances.

Malone and Brünken (2013) provide empirical evidence for this assumption. They assessed car driving related knowledge in an expert-novice comparison and applied either useful animations or static pictures to visualize the same dynamic processes in traffic scenarios. They found an interaction effect between presentation mode (static vs. dynamic) and expertise. Experts outperformed novices only in the static version of the test. The animations were helpful for the novices, as they were relieved from the need to infer motion from a static picture. In contrast, the expert drivers did not benefit from the presentation of animations because, based on their experience, they were able to mentally animate the static pictures, easily. The authors showed in their experiment how introducing helpful features in assessment (for example by providing animations) could interfere with criterion validity.

Another study by Brünken, Steinbacher, Schnotz, and Leutner (2001) also provides evidence that CBA efforts, such as using codality, have to be considered in the frame of CBMMA to guarantee the required validity of assessments. They showed that effects of learning can be detected more easily when posttest materials are presented in the same codality as learning materials. It depends on the learning goal if this really is valid assessment. Two other studies (Brünken, Plass, & Leutner, 2004; Brünken, Steinbacher, Plass, & Leutner, 2002) show the same effects for valid and reliable measures of cognitive load when considering the modality principle in the frame of using dual-task methods. The used modality should be the same in both, the CBMMA and the previously presented computer-based multimedia material that often also includes narration instead of text (i.e., audio-files). In other words, dual-task methods appear to be modality-specific, at least when using visual or auditory prompts within the dual-task method for measuring cognitive load. And this specificity can be used in an advantageous way to filter out the corresponding interesting cognitive processes.

As summarized by Park and Brünken (2015), within the dual-task paradigm, cognitive load is measured by the performance of a secondary task executed parallel to / simultaneously with the primary learning task. In detail, the dual-task method measures cognitive load at different times of measurement during learning (primary task) with the help of the secondary task performance (e.g., reaction time to a signal), which reflects the amount of cognitive load in the primary task. In other words, differences in a learner’s resource consumption caused by different presentations of the learning material can be measured by differences in performance on the secondary task. The established secondary tasks usually include either an auditory or visual cue in the instruction. For example, Brünken et al. (2004) asked participants to monitor a letter in the upper part of the computer screen and react by pressing the space bar when a color change was observed. In a recent study by Park and Brünken (2015) using a continuous, intra-individual and behavioral measure the new task is achieved by utilizing internalized cues. More specifically, a previously practiced rhythm is executed continuously by foot tapping (i.e., the secondary task) while learning (i.e., primary task). Execution precision was used as indicator for cognitive load; the greater the precision, the lower the load. This is a variation of dual-task may provide a general indicator for cognitive load in that it is not modality-specific for executive control processes (Baddeley, 1992), but this needs further empirical testing.

It is likely, thus, that experts outperform novices only on those tasks that match their area of expertise (Chi, 2006). To best detect the specific level of expertise, one has to find a 'standardized set of tasks' that are most 'representative' for a domain (Ericsson & Lehmann, 1996; Ericsson & Smith, 1991). For few domains these tasks are static and greyscale and thus easily presentable on paper (even though, most of expertise research has been conducted in such tasks: Reingold & Sheridan, 2011). Only recently, expertise is investigated in more authentic, and thus ecologically valid tasks and also capturing the relevant underlying processes, for instance with eye tracking (e.g., Balslev et al., 2012; Jaarsma, Jarodzka, Nap, Van Merriënboer, & Boshuizen, 2015; Jarodzka, Scheiter, Gerjets, & Van Gog, 2010; Van Meeuwen et al., 2014; Wolff, Jarodzka, Van den Bogert, & Boshuizen, 2016). Only such research can unravel the exact processes underlying different levels of expertise and thus, ultimately allow for their assessment and prediction.

Using More Sophisticated Response Modes in Multimedia Assessment

Theoretical assumptions and empirical evidence do not support the uncritical transfer of design guidelines for multimedia learning to multimedia assessment. Processes of expertise development can on the one side explain why the transfer won't work and on the other hand they point to possible resolutions for the problem.

Van Gog, Ericsson, Rikers, and Paas (2005) have already addressed a part of the problem in their theoretical paper on the need for special guidelines to design instructional materials for advanced learners. The authors discuss, why many design principles that work for novice learners might be inappropriate for advanced learners (expertise reversal effect) and that there is a need for special instructional design guidelines for learners which already have gained prior knowledge and made experiences in a domain. In order to do that, Van Gog et al. (2005) advise to take research findings on expert-novice differences, expertise acquisition and factors that have proven to foster expertise into account. The authors emphasize the need for appropriate knowledge and skill assessment to be able to design adapted instruction.

In their literature review, Ericsson, Krampe, and Tesch-Römer (1993) found that while practice is essential for skill and expertise development, whether performance is maximized by practice depends on how something is practiced. Particularly, the amount of *deliberate practice* is crucial for expertise development. Deliberate practice involves the explicit aim to improve one's skill, permanent effort, phases of direct instruction, and immediate feedback. Expertise assessment should reveal if a person has engaged in deliberate practice and has acquired correct schemata. Expert performance is defined by Ericsson and Lehman (1996) as "an extreme adaptation to task constraints" (p. 291). The assessment of expertise therefore requires the selection of the essential aspects of expert performance, the identification of relevant real task constraints and the creation of representative tasks for the specific domain. This approach ensures ecological as well as criterion validity of the assessment.

For complex dynamic domains, such as many sports, these two validity criteria, ecological and criterion validity, are considered to be related. As Hodges, Huys, and Starkes (2007) report, the increase of ecological validity of stimulus and response modes of the tasks (facilitated by the means of new media) makes them more sensitive to expert-novice differences, which indicate criterion validity. This finding leads to the conclusion that valid tests need to include tasks that are representative for a certain domain. Representative tasks (task demands match the requirements imposed usually in the specific domain) can be created by the means of multimedia assessment by the application of sophisticated response modes (e.g., reaction time measure; car/truck/flight simulation).

However, it is not only relevant *what type* of response mode we use for assessment, but also how the assessees *processes* them. Research focusing on the processes underlying multimedia assessment indicates that two issues are crucial (Jarodzka, Janssen, Kirschner, & Erkens, 2015; Ögren, Nyström, & Jarodzka, 2016): First, the students must carefully process the main question posed to them. Second, they must integrate this question with the multimedia material (e.g., in forms of integrative saccades). Only such a processing behavior could be related to higher assessment scores.

Conclusion

We conclude this chapter with a number of general considerations with respect to the application of CTMMA and an elucidation (see Table 1) of the similarities and differences between CTMML and CTMMA.

From the perspective of cognitive load research, at first, intrinsic cognitive load can be considered in the given assessment tasks by varying the complexity. This is already considered in several studies by using for example retention versus comprehension and transfer tasks (e.g. Marcus et al, 1996) or using tasks asking for learning outcomes of processes in contrast to knowledge about structures (e.g. Park et al., 2016). Second, extraneous cognitive load can be considered by varying this type of load to ensure ecological validity (i.e., increase extraneous cognitive load) and assess the limits of the learner (i.e., present different tasks with increasing extraneous cognitive load levels). Integrating at this point also the perspective of research on multimedia learning, optimal extraneous cognitive load should be imposed by 'ignoring' many of the instructional design principles given to reduce extraneous cognitive load. However, this should be operationalized carefully, not overdoing it. In addition, response modes have to be considered, as these allow for representative tasks by means of employing sophisticated response modes (i.e., authenticity measures are needed). Third, when considering germane cognitive load, the learner's engagement in the assessment situation should be varied in order to test the limits of the learner. In Table 1 some concluding CTMMA principles are summarized hinting at the concrete possible operationalizations for considering CTMMA in future assessment.

Moreover, important general issues that have to be kept in mind when designing multimedia assessment are the aims/goals of assessment and level of expertise of the person/group to be assessed, the content of the assessment tasks and the type of knowledge and skills that the assessment is intended to capture, as well as the fact that the design of the assessment tasks on the computer screen is quite different from doing this on paper.

Table 1.

Principle	*CTMML*	*CTMMA*
Modality	People learn better from graphics and narrations than from animation and on-screen text, especially when the graphic is complex, the words is familiar, and the lesson is fast-paced.	Pictures can actually 'trick' assessees into confirming a statement (cf. Ögren, Nyström, & Jarodzka, 2016). Hence, they should be used scarcely and cautiously.
Segmentation	People learn better from a multimedia lesson is presented in user-paced segments rather than as a continuous unit.	It is easier to distinguish between individuals of higher and lower expertise, if the task or problem is presented as a continuous unit (cf. whole task).
Pre-training	People learn better from a multimedia lesson when they know the names and characteristics of the main concepts.	It is easier to distinguish between individuals of higher and lower expertise, if no pre-training on the test-material was given.
Coherence	People learn better when extraneous words, pictures and sounds are excluded rather than included. Adding interesting but irrelevant materials to e-learning courses may distract the learner.	It is easier to distinguish between individuals of higher and lower expertise, if the amount of coherence of the testing material corresponds to the coherence found in the real-world.
Signaling	People learn better when cues that highlight the organization of the essential material are added.	It is easier to distinguish between individuals of higher and lower expertise, if no additional cues or highlights are given.
Spatial contiguity	People learn better when corresponding words and pictures are presented near rather than far from each other on the page or screen.	It is easier to distinguish between individuals of higher and lower expertise, if the spatial contiguity of the given material corresponds to the real-world situation: the assessment itself is that the assessees select and integrate the relevant information autonomously.
Temporal contiguity	People learn better when corresponding words and pictures are presented simultaneously rather than successively	It is easier to distinguish between individuals of higher and lower expertise, if all information is presented in such a way, as it would occur in the real-world task: for some situations this may mean that people need to integrate a lot of information at the same time or that they need to remember information for later usage.
Redundancy	People learn better from graphics and narration than from graphics, narration and on-screen text. The visual text information presented simultaneously to the verbal information becomes redundant.	It is easier to distinguish between individuals of higher and lower expertise, if the amount of redundant information is as high as it would be in the according real-world task or problem.
Emotional Design	People learn better from multimedia lessons when	Emotional design, emotion induction or

and Emotion Induction including Personalization and Voice	words are in conversational style rather than formal style. People learn better when the narration in multimedia lessons is spoken in a friendly human voice rather than a machine voice.	personalization and the use of human voice within the assessment situation could help to distinguish between individuals of higher and lower expertise because experts are known to be capable to compensate effects of emotionalized material, induced emotions or formal instead of conversational style.
Self-pacing	Learners learn better from self-paced than from system-paced multimedia lessons	As experts are known to execute tasks faster than novices do, putting temporal restrictions to assessment (presentation and answer time) may help easier distinguish between individuals of higher and lower expertise.

Acknowledgements

This work was - in part - supported by the German Federal Ministry of Education and Research (01PL12057).

Author Information

Complete name: Kirschner, Paul A.

Institutional affiliation: Open University of the Netherlands & University of Oulu, Finland

Institutional address: Valkenburgerweg 177; 6419 AT Heerlen; The Netherlands

Complete mailing address: Paul Kirschner; Valkenburgerweg 177; 6419 AT Heerlen; The Netherlands

Telephone number: +31 (0)45 5762361

Email address: paul.kirschner@ou.nl

Website http://www.ou.nl/web/pki

Short biographical sketch: Paul A. Kirschner is Distinguished University Professor at the Open University of the Netherlands as well as Visiting Professor of Education with a special emphasis on Learning and Interaction in Teacher Education at the University of Oulu, Finland. He was previously scientific director of the Learning and Cognition program at CELSTEC, Opn University of the Netherlands. He is an internationally recognized expert in the fields of educational psychology and instructional design. He is a past President (2010-2011) of the International Society for the Learning Sciences and a former member of the Dutch Educational Council and, as such, was advisor to the Minister of Education (2000-2004). He is also a member of the Scientific Technical Council of the Foundation for University Computing Facilities (SURF WTR), chief editor of *Journal of Computer Assisted Learning* and associate editor of C*omputers in Human Behavior*. As for books, he is co-author of the recently released book *Urban Myths about Learning and Education* as well as of the highly successful book *Ten steps to complex* *learning*, and editor of two other books (*Visualizing Argumentation* and *What we know about CSCL*).

Complete name: Park, Babette
Institutional affiliation: Saarland University
Institutional address: Campus A4 2, 66123 Saarbrücken, Germany
Complete mailing address: Saarland University, Jun.-Prof. Dr. Babette Park, Campus A4 2, D-66123 Saarbrücken, Germany
Telephone number: +49 681 302 2340
Email address: b.park@mx.uni-saarland.de
Short biographical sketch: Babette Park is assistant professor (Juniorprofessorin) at the Department of Education at Saarland University in Germany. She has a Diplom in Psychology and finished her German PhD in Educational Psychology in 2010. Since 2012 she holds the chair "Empirical Research and Didactics in Higher Education" funded by the German Federal Ministry of Education and Research (01PL12057) within the project "Studying with Profile: Competence in Research and Practice". Babette Park teaches students of Psychology (Diplom/Bachelor/Master), of the international and interdisciplinary study course Educational Technology (Master), and of Teacher Education. Additionally, she is teacher and counselor for academic staff at the Centre for Key Skills and Didactics in Higher Education. Her research group runs three different research lines. Within basic-oriented research, suitable instruments are developed and validated for measuring cognitive load. With instructional-psychological research, cognitive and affective processes are investigated in multimedia learning. Finally, application-oriented research focuses on analyzing conditions and determinants of successful teaching in higher education from

an educational-psychological perspective. Babette Park is an active member of an international network in research on cognitive load and multimedia learning. Since 2015 she is member of the Cognitive Load Theory Advisory Committee for the annually meeting International Cognitive Load Theory Conference.

Complete name: Malone, Sarah
Institutional affiliation: Saarland University
Institutional address: Campus A4 2, 66123 Saarbrücken, Germany
Complete mailing address: Sarah Malone, Campus A4 2, 66123 Saarbrücken, Germany
Telephone number: +49 681 302 58336
Email address: s.malone@mx.uni-saarland.de
Website: http://www.uni-saarland.de/lehrstuhl/bruenken/personen/dr-sarah-malone.html
Short biographical sketch: Sarah Malone works as a postdoctoral research fellow at the Department of Education, Saarland University, Germany. Her research focuses on learning and assessment with multiple representations in mathematics and traffic psychology (Hazard Perception assessment, computer based driver training).

Complete name: Jarodzka, Halszka
Institutional affiliation: Open University of the Netherlands & Lund University, Sweden
Institutional address: Valkenburgerweg 177; 6419 AT Heerlen; The Netherlands

Complete mailing address: Halszka Jarodzka; Valkenburgerweg 177; 6419 AT Heerlen; The Netherlands

Telephone number: +31 45 576 2410

Email address: Halszka.Jarodzka@OU.nl

Website: http://www.ou.nl/web/hjr

Short biographical sketch: Halszka Jarodzka works as an assistant professor at the Open University of the Netherlands, where she is chair of the topic group "Processes of learning and expertise development in information-rich environments". Moreover, she works part-time as a visiting scholar at a large eye tracking laboratory at Lund University in Sweden. Her research deals with the use of eye tracking to understand and improve learning and its instruction. In that, she investigates three topics: first, the instructional design of computer-based learning and testing environments; second, the characteristic and development of visual expertise in diverse professions; and third, training of perceptual skills with eye movement modeling examples (EMME). She is co-founder and chair of a special interest group on "Online measures of learning" within the European Association of Research on Learning and Instruction.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89-192). New York: Academic Press.

Baddeley, A. D. (1992). Working memory. *Science, 255*, 556-559. doi: 10.1126/science.1736359

Balslev, T., Jarodzka, H., Holmqvist, K., De Grave, W. S., Muijtjens, A., Eika, B., . . . Scherpbier, A. (2012). Visual expertise in paediatric neurology. *European Journal of Paediatric Neurology, 16*, 161-166. doi: 10.1016/j.ejpn.2011.07.004

Bortz, J., & Döring, N. (2013). *Forschungsmethoden und evaluation* [Research methods and evaluation]. Heidelberg, Germany: Springer-Verlag.

Boshuizen, H. P. A., & Schmidt, H. G. (1992). Biomedical knowledge and clinical expertise. *Cognitive Science, 16*, 153-184.

Brünken, R., Plass, J. L., & Leutner, D. (2004). Assessment of cognitive load in multimedia learning with dual-task methodology: Auditory load and modality effects. *Instructional Science, 32*, 115-132. doi: 10.1023/B:TRUC.0000021812.96911.c5

Brünken, R., Seufert, T., & Paas, F. (2010). Measuring cognitive load. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 181-202). Cambridge: University Press.

Brünken, R., Steinbacher, S., Plass, J., & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology*, *49*, 109-119. doi: 10.1027//1618-3169.49.2.109

Brünken, R., Steinbacher, S., Schnotz, W., & Leutner, D. (2001). Mentale Modelle und Effekte der Präsentations- und Abrufkodalität beim Lernen mit Multimedia [Mental models and effects of presentation and retrieval coding when learning with multimedia]. *Zeitschrift für Pädagogische Psychologie, 15*, 16-27. doi: 10.1024//1010-0652.15.1.16

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and instruction, 8*, 293-332. doi: 10.1207/s1532690xci0804_2

Chi, M. T. H. (2006). Two approaches to the study of experts' characteristics. In K. A. Ericsson, N. Charness, R. R. Hoffman & P. Feltovich (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 21-30). Cambridge: Cambridge University Press.

De Boer, N. (2009). *De computer bij de centrale examens. Duidelijk digitaal 2* [The computer at the national exams. Clearly digital 2]. http://www.cito.nl/VO/ce/compex/introductie/cve_comp_bij_ce_duidelijk_digitaal_2.pdf

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100*, 223-234. doi: 10.1037/0022-0663.100.1.223

Dennick, R., Wilkinson, S., & Purcell, N. (2009). Online eAssessment: AMEE guide no. 39. *Medical teacher, 31*, 192-206. doi: 10.1080/01421590902792406

Ericsson, K. A., Charness, N., Hoffman, R. R., & Feltovich, P. (Eds.). (2006). *The Cambbridge handbook of expertise and expert performance*. Cambridge: University Press.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, 363-406.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology, 47*, 273–305.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaption to task constraints. . *Annual Reviews in Psychology, 47*, 273-305.

Ericsson, K. A., & Smith, J. (1991). Prospects and limits in the empirical study of expertise. In K. A. Ericsson & J. Smith (Eds.), *Towards a general theory of expertise: Prospects and limits* (pp. 1-38). Cambridge, MA: University Press.

Fischer, F., Waibel, M., & Wecker, C. (2005). Nutzenorientierte Grundlagenforschung im Bildungsbereich [Benefit-oriented basic research in the field of education]. *Zeitschrift für Erziehungswissenschaft, 8*, 427-442. doi: 10.1007/s11618-005-0149-7

Hartig, J., & Klieme, E. (Eds.). (2007). *Möglichkeiten und Vorraussetzungen technologiebasierter Kompetenzdiagnostik* [Possibilities and prerequisites of technology-driven competence diagnostics]. Bonn, Berlin: Bundesministerium für Bildung und Forschung (BMBF).

Hamm, S., & Robertson, I. (2010). Preferences for deep-surface learning: A vocational education case study using a multimedia assessment activity. *Australasian Journal of Educational Technology, 26*, 951-965. doi: 10.14742/ajet.1027

Hodges, N. J., Huys, R., & Starkes, J. L. (2007). Methodological review and evaluation of research in expert performance in sport. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of Sport Psychology* (Vol. 3, pp. 161-183). Hoboken, NJ: John Wiley & Sons.

Jaarsma, T., Jarodzka, H., Nap, M., Van Merriënboer, J. J. G., & Boshuizen, H. P. A. (2015). Expertise in clinical pathology: Bridging the gap. *Advances in Health Sciences Education, 20*, 1089-1106. doi: 10.1007/s10459-015-9589-x

Jarodzka, H., Janssen, N., Kirschner, P. A., & Erkens, G. (2015). Avoiding split attention in computer-based testing: Is neglecting additional information facilitative? *British Journal of Educational Technology, 46*, 803-817. doi: 10.1111/bjet.12174

Jarodzka, H., Scheiter, K., Gerjets, P., & Van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Journal of Learning and Instruction, 20*, 146-154. doi: 10.1016/j.learninstruc.2009.02.019

Jurecka, A., & Hartig, J. (2007). Computer- und netzwerkbasiertes Assessment [Computer- and network-based assessment]. In J. Hartig & E. Klieme (Eds.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (pp. 37-48). Bonn, Berlin: Bundesministerium für Bildung und Forschung (BMBF).

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*, 23-32. doi: 10.1207/S15326985EP3801_4

Kirschner, P. A. (2002). Cognitive load theory: Implications of cognitive load theory on the design of learning. *Learning and Instruction, 12*, 1-10. doi: 10.1007/s11251-009-9110-0

Knörzer, L., Brünken, R., & Park, B. (2016). Facilitators or suppressors: Effects of experimentally induced emotions on multimedia learning. *Learning & Instruction, 44*, 97-107. doi: 10.1016/j.learninstruc.2016.04.002

Korbach, A., Brünken, R., & Park, B. (2016). Learner characteristics and information processing in multimedia learning: A moderated mediation of the seductive details effect. *Learning and Individual Differences, 51*, 59-68. doi:10.1016/j.lindif.2016.08.030

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

Lienert, G. A. (1969). *Testaufbau und Testanalyse* [Test Construction and Test Analysis] (3., durch einen Anh. über Faktorenanalyse erg. Aufl.) . Weinheim: Beltz.

Lienert, G. A., & Raatz, U. (1994). *Testaufbau und Testanalyse* [Test Construction and Test Analysis] (5. völlig neu bearbeitete und erweiterte Auflage). Weinheim: Beltz.

Malone, S., & Brünken, R. (2013). Assessment of driving expertise using multiple choice questions including static vs. animated presentation of driving scenarios. *Accident Analysis & Prevention, 51*, 112-119.

Marcus, N., Cooper, M., & Sweller, J. (1996). Understanding instructions. *Journal of Educational Psychology*, *88*, 49-63.

Marshall, S. P. (2002). The Index of Cognitive Activity: Measuring cognitive workload. Proceeding of the 2002 IEEE 7th Conference, *Human Factors and Power Plants, 2002*. doi: 10.1109/HFPP.2002.1042860

Mayer, R. E. (1996). Learning strategies for making sense out of expository text: The SOI model for guiding three cognitive processes in knowledge construction. *Educational Psychology Review, 8*, 357-371.

Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.

Mayer, R. E. (2005a). *The Cambridge handbook of multimedia learning*. New York: Cambridge University Press.

Mayer, R. E. (2005b). Cognitive theory of multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). New York: Cambridge University Press.

Mayer, R. E. (2009). *Multimedia learning*. Cambridge: University press.

Mayer, R. E. (2010). Unique contributions of eye-tracking research to the study of learning with graphics. *Learning and Instruction, 20*, 167-171. doi: doi:10.1016/j.learninstruc.2009.02.012

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

Meyer, C. A. (1992). What's the Difference between Authentic and Performance Assessment? *Educational Leadership, 49*, 39-40.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Miyake, A., & Shah, P. (Eds.) (1999). *Models of working memory: Mechanisms of active maintenance and executive control.* New York: Cambridge University Press.

Moosbrugger, H., & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion* [Test theory and questionnaire design]. Berlin, Germany: Springer.

Moreno, R., & Park, B. (2010). Cognitive load theory: Historical development and relation to other theories. In R. M. R. B. J. L. Plass (Ed.), *Cognitive load theory* (pp. 9-28). New York, USA: Cambridge University Press.

Münzer, S. (2012). Facilitating spatial perspective taking through animation: Evidence from an aptitude-treatment-interaction. *Learning and Individual Differences, 22*, 505-510. http://dx.doi.org/10.1016/j.lindif.2012.03.002.

Münzer, S. (2015). Facilitating recognition of spatial structures through animation and the role of mental rotation ability. Learning and *Individual Differences, 38*, 76-88. http://dx.doi.org/10.1016/j.lindif.2014.12.007.

Münzer, S., Seufert, T., & Brünken, R. (2009). Learning from multimedia presentations: Facilitation function of animations and spatial abilities. *Learning and Individual Differences, 19*, 481-485. doi: 10.1016/j.lindif.2009.05.001

Ögren, M., Nyström, M., & Jarodzka, H. (2016). There's more to the multimedia effect than meets the eye: Is seeing pictures believing? *Instructional Science*. doi: 10.1007/s11251-016-9397-6

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

Paivio, A. (1986). *Mental representations: A dual coding approach*. New York: Oxford University Press.

Park, B. (2010). Testing the additivity hypothesis of cognitive load theory (Doctoral dissertation, Saarland University, Saarbrücken, Germany). Retrieved from http://scidok.sulb.uni-saarland.de/volltexte/2010/3478/.

Park, B., & Brünken, R. (2015). The rhythm method: A new method for measuring cognitive load: An experimental dual-task study. *Applied Cognitive Psychology, 29*, 232-243. doi: 10.1002/acp.3100

Park, B., Flowerday, T., & Brünken, R. (2015). Cognitive and affective effects of seductive details in multimedia learning. *Computers in Human Behavior, 44*, 267-278. doi: 10.1016/j.chb.2014.10.061

Park, B., Knörzer, L., Plass, J. L., & Brünken, R. (2015). Emotional design and positive emotions in multimedia learning: An eyetracking study on the use of antropomorphisms. *Computers and Education, 86*, 30-42. doi: 10.1016/j.compedu.2015.02.016

Park, B., Korbach, A., & Brünken, R. (2015). Do learner characteristics moderate the seductive-details-effect? A cognitive-load-study using eye-tracking. *Journal of Educational Technology & Society, 18*, 24-36. http://www.ifets.info/journals/18_4/3.pdf, Creative Commons CC-BY-ND-NC 3.0

Park, B., Münzer, S., Seufert, T., & Brünken, R. (2016). The role of spatial ability when fostering mental animation in multimedia learning: An ATI-study. *Computers in Human Behavior, 64*, 497-506. doi: 10.1016/j.chb.2016.07.022

Plass, J. L., Moreno, R., & Brünken, R. (2010). *Cognitive load theory*. New York: Cambridge University Press.

Posner, M. I. (1988). Introduction: What is it to be an expert? In M. T. H. Chi, R. Glaser & M. J. Farr (Eds.), *The nature of expertise*. Hillsdale, NJ: Erlbaum.

Reingold, E. M., & Sheridan, H. (2011). Eye movements and visual expertise in chess and medicine. In S. P. Liversedge, I. D. Gilchrist & S. Everling (Eds.), *Oxford handbook of eye movements* (pp. 523-550). Oxford, UK: Oxford University Press.

Rumelhart, D. E., & Ortony, A. (1976). *The representation of knowledge in memory*: Center for Human Information Processing, Department of Psychology, University of California, San Diego.

Schank, R., & Abelson, R. (1977). Scripts, goals, and understanding. *Hillsdale, NJ: LEA. Sharkey, NE, & Mitchell, DC (1985). Word recognition in a functional context: The use of scripts in reading. Journal of Memory and Language, 24*, 253-270.

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

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychological Review, 10*, 251-296. doi: 10.1023/b:truc.0000021808.72598.4d

The International Test Commission (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing., 6*, 143-171. doi: 10.1207/s15327574ijt0602_4

van Gog, T., Ericsson, K. A., Rikers, R. M. J. P., & Paas, F. (2005). Instructional design for advanced learners: Establishing connections between the theoretical frameworks of cognitive load and deliberate practice. *Educational Technology Research and Development, 53*, 73-81. doi: 10.1007/BF02504799

Van Meeuwen, L. W., Jarodzka, H., Brand-Gruwel, S., Kirschner, P. A., De Bock, J. J. P. R., & Van Merriënboer, J. J. G. (2014). Identification of effective visual problem solving strategies in a complex visual domain. *Learning and Instruction, 32*, 10-21. doi: 10.1016/j.learninstruc.2014.01.004

Van Mierlo, C. M., Jarodzka, H., Kirschner, F., & Kirschner, P. A. (2012). Cognitive load theory and e-learning. In Z. Yan (Ed.), *Encyclopedia of cyber behavior*. Hershey, PA: IGI Global.

Wolff, C. E., Jarodzka, H., Van den Bogert, N., & Boshuizen, H. P. A. (in press). Teacher vision: Comparing expert and novice teachers' perception of problematic classroom management scenes. *Instructional Science*.