Taking a Closer Look at the Desirability of Desirable Difficulties— Additionally Focusing on Required Prerequisites, Negative Side-Effects, and Negative Consequences

Kumulative Dissertation

zur Erlangung des akademischen Grades Doktorin der Philosophie (Dr. phil.)

Vorgelegt im Fachbereich 01 Humanwissenschaften

der Universität Kassel

Von M.Sc. Kristin Wenzel

< Datum Einreichung der Dissertation: 15.12.2021 >

< Datum der Disputation: 20.05.2022 >

< Kassel >

Dekanin des Fachbereichs Humanwissenschaften:	Prof. Dr. Natalie Fischer
Betreuer:	Prof. Dr. Marc-André Reinhard
Erstgutachter:	Prof. Dr. Ralf Rummer
Zweitgutachterin:	Prof. Dr. Mirjam Ebersbach
Weitere Mitglieder der Promotionskommission:	Prof. Dr. Judith Schweppe
	Prof. Dr. Martin Hänze

Contents

Overview

As the title of my dissertation already indicates, the main purpose of the following work was to take a closer, more thorough, and encompassing look at the desirability of *desirable difficulties* (e.g., R. A. Bjork, 1994). The term desirable difficulties thereby subsumes varying types of difficult, demanding, and challenging learning strategies—like the application of (learning/practice) tests and generation tasks: Working on and successfully solving those difficult tasks requires more (cognitive) effort compared to working on easier learning tasks but thereby leads to increased learning outcomes after a delay (e.g., R. A. Bjork, 1994; Bjork & Bjork, 2020). Due to these beneficial long-term learning effects, such difficulties are called desirable. However, previous literature also showed that special prerequisites have to be given (e.g., higher previous knowledge or higher reading abilities) for learners to be even able to successfully overcome difficulties and to be actually able to reap their benefits (see e.g., McDaniel et al., 2002; McNamara et al., 1996). Hence, it is possible that not all learners benefit equally from such effortful learning tasks and that they might not be desirable for everyone. Additionally, previous work also found that learners often perceive difficult tasks as negative and stress inducing (see e.g., Hinze & Rapp, 2014; Khanna, 2015; O'Neil et al., 1969), thus implying that the usage of desirable difficulties might be rather unpleasant for learners. Such negative side-effects were then, in turn, often linked to increased (academic) cheating as well as to impaired later learning outcomes and a reduced effectiveness of the applied strategies (e.g., Brimble & Stevenson-Clarke, 2005; Seipp, 1991; Whitley, 1998; Wowra, 2007). Hence, desirable difficulties might lead to negative side-effects directly during learning and might also cause even more negative consequences later on. These findings and assumptions indicate that further dimensions or further factors determining and influencing the (und)desirability of desirable difficulties should be considered in addition to their effects on long-term learning. Therefore, this dissertation

focused on potential prerequisites determining learners' (un-)favourable views of desirable difficulties and the benefits learners are able to reap through using desirable difficulties. Moreover, I also focused on negative side-effects and on further negative consequences that desirable difficulties may directly and indirectly cause. These important—and often related—issues must be thoroughly addressed before it is possible to give recommendations if, how, or for whom desirable difficulties can be applied. In addition, my dissertation should also be stimulating for further research, insofar as that it highlights relevant research issues that need to be addressed, that it presents open questions that require more empirical testing, and that it discusses potential implications for future work.

To present the just described issues as best as possible, the following work consists of two parts: a synopsis and an empirical section. The synopsis thereby serves as a frame for the published research papers embedded in the empirical section (see Appendix A to F). Thus, the synopsis links short summarizations and brief descriptions of my conducted work to already existing literature and discusses findings, implications, recommendations, and future directions of my research. The empirical section then presents all studies that build this dissertation in paper-based formats—hence, making my papers available for in-depth reading to expand the relatively few information given in the synopsis.

The synopsis starts with an introduction of different types of desirable difficulties and their underlying theoretical basis. My dissertation—and my conducted studies—thereby focusses mostly on the application of tests as a common and extremely effective desirable difficulty. I then describe cognitive-motivational learner characteristics that might serve as prerequisites for positive attitudes towards desirable difficulties, for learners' usage of difficult tasks, and for the benefits elicited by desirable difficulties. I then concentrate more strongly on cognitive abilities that might act as prerequisites or boundary conditions for the long-term learning effects of desirable difficulties. More specifically, I introduce previous and own work investigating if learners benefit differently from the application of difficulties

depending on their intelligence. Subsequent, the synopsis further contemplates how learners experience the application of tests and if—and to what extent—tests lead to negative side-effects like more negative evaluations and increased stress perceptions during learning. I then describe theoretical assumptions and empirical findings indicating that tests might further also directly and indirectly cause negative consequences like increased academic cheating in later examinations. Thereafter, my synopsis will contemplate reduced later learning outcomes and impaired effectiveness of tests as further potential negative consequence indirectly caused by the application of such tests. I thereby also describe effects of learners' intelligence and highlight linkages among the different research issues of my dissertation (concerning prerequisites, negative side-effects, and negative consequences). At the end of the synopsis, I discuss, among others, for whom tests are effective, how tests could be implemented to be beneficial for every learner, and how their benefits can be reaped without suffering under potential negative side-effects or negative consequences.

In the following empirical section, the 8 studies embedded in my 6 paper-based manuscripts will be presented following the order in which they were first mentioned in the synopsis. The first paper thereby consists of one online study addressing linkages among cognitive-motivational characteristics (internal locus of control, self-efficacy, and trait stress), learners' attitudes towards desirable difficulties, and their self-reported usage of those (see Appendix A). The next paper then presents a classroom study exploring the influence of learners' performance expectancies on long-term learning effects of generation tasks (see Appendix B). After that, the third paper summarized in my synopsis tested in two laboratory studies if intelligence serves as a cognitive prerequisite for the benefits of desirable difficulties (see Appendix C): The first study thereby found no long-term learning effects of generation tasks, whereas the second study resulted in a beneficial effect of a test compared to a re-reading task. This effect was, in turn, moderated by intelligence and showed that only at least averagely intelligent learners profited from taking tests. The next paper then investigated

how learners perceive and experience the application of tests and if tests lead to negative side-effects during learning (see Appendix D): The findings of one online and one laboratory study showed that tests caused more negative evaluations of the learning situation, higher immediate stress perceptions, and higher acute anxiety experiences. My fifth paper then focused on the assumption that tests directly and indirectly lead to more academic cheating (see Appendix E): The conducted online study yielded that hypothetical learning situations including tests indirectly (via heightened negative evaluations of the situation) increased likelihoods of cheating and justifications for cheating in a later hypothetical examination. Finally, the sixth and last paper included in my empirical section simultaneously focused on benefits of tests, on negative side-effects triggered by tests, on further caused negative consequences, and on learners' intelligence (see Appendix G): The respective laboratory study showed that a short test was more beneficial for later learning than re-reading, but that it also caused more acute stress perceptions, which in turn suppressed the beneficial effects of tests. Notably, learners' intelligence was negatively correlated to their stress perceptions and positively correlated to their later learning outcomes—but did not moderate any of the other effects.

SYNOPSIS

Desirable Difficulties and Their Beneficial Long-Term Learning Effects

If asked, students, lecturers, and researchers would most likely all agree that durable long-term learning, successful knowledge acquisition, and increased academic achievement are the most desired outcomes of learning strategies applied in schools and universities. However, although they would agree regarding preferred learning outcomes, students and lecturers nonetheless often disagree with researchers regarding the best ways to achieve these goals. For instance, when choosing or evaluating learning strategies, students often use their immediate subjective experiences while learning as cues for the effectiveness of the learning task—hence, students (and even their lecturers) often prefer and apply strategies that feel easy, fluent, and good (like re-reading; e.g., Bjork & Bjork, 2019; R. A. Bjork et al., 2013; Karpicke et al., 2009; Kornell et al., 2011; Rivers, 2021). Students and lecturers also typically think that efficient learning is easy learning and that information that is easily processed, encoded, or retrieved is information that has already been learned well (e.g., Biwer et al., 2020; Bjork & Bjork, 2019; R. A. Bjork et al., 2013; Kornell et al., 2011; Rivers, 2021). Accordingly, most students and lecturers believe that specific learning strategies that most fit students' learning styles or their (cognitive) abilities should require little effort (e.g., Bjork & Bjork, 2019; R. A. Bjork et al., 2013). In contrast to these assumptions, researchers have often argued that easy learning strategies are rather ineffective in the long run, whereas more difficult, more demanding, and more effortful learning strategies are especially beneficial for durable long-term learning outcomes (e.g., Bjork & Bjork, 1992, 2011, 2019; Diemand-Yauman et al., 2011; Dobson & Linderholm, 2015; Karpicke et al., 2009; Kornell et al., 2011). This applies to the afore-mentioned desirable difficulties (e.g., R. A. Bjork, 1994; Bjork & Bjork, 2020), which include varying intentionally hindered learning tasks that require considerable but still manageable cognitive effort during learning. Successfully solving or

overcoming such difficulties is not immediately beneficial but—more important—elicits desirable cognitive processes that strengthen memory and increase delayed long-term learning outcomes (e.g., Bjork & Bjork, 1992, 2011, 2019). Because these findings contrast with students and lecturers' typical beliefs and (mis-)conceptions regarding the effectiveness of learning strategies, it is important to clearly communicate which learning tasks are beneficial in the long run, to thoroughly describe potentially required prerequisites or boundary conditions of desirable difficulties, and to highlight possible negative side-effects or negative consequences caused by such difficult learning tasks.

Desirable difficulties include, among others, *distributed practice* (e.g., using spaced instead of massed learning episodes; e.g., Cepeda et al., 2006; Ebersbach & Barzagar Nazari, 2020; C. E. Greving & Richter, 2021), *disfluency* (e.g., interrupting fluency through harder-to-read fonts; e.g., Diemand-Yauman et al., 2011; Eitel et al., 2014; Weissgerber & Reinhard, 2017), and *interleaving* (e.g., mixing different learning topics; e.g., Brunmair & Richter, 2019; Nemeth et al., 2021; Ziegler & Stern, 2014). Two especially robust, easily applicable, and empirically well-documented types of desirable difficulties are *generation tasks* and (*learning/practice*) *tests*: In the beginning, my dissertation will thereby take a closer look on both these related types of desirable difficulties—however, the main focus will be on the application of tests. The literature on generation tasks (also often known as: *generation, generation effect,* or *problem-solving*) has often shown that active (self-)generation (of e.g., answers, solutions, questions, or examples) is more beneficial than passive consummation of materials through re-reading texts, studying already solved problems, or memorizing already worked examples (e.g., Bertsch et al., 2007; McCurdy et al., 2020). These beneficial long-term learning effects of generation tasks were obtained in varying learning settings (e.g., schools, universities, or laboratories), for younger and older students, for different forms of generation tasks (e.g., completing word fragments or sentences, creating synonyms, filling blanks, generating examples or questions, completing exercises, or working on problem-

solving tasks), and for a wide range of naturalistic, complex, and curricular topics (e.g., astronomy, mathematics, or physics; e.g., Bertsch et al., 2007; Ebersbach et al., 2020; McCurdy et al., 2020; Moreno et al., 2009; Richland et al., 2005). Regarding the application of tests (also often known as: *quizzing, retrieval practice, testing, testing effect*, or *test enhanced learning*), previous work has also repeatedly shown that (learning/practice) tests or quizzes conducted after an initial study opportunity strongly benefit learners' later learning outcomes (e.g., Adesope et al., 2017; Dobson & Linderholm, 2015; Pan & Rickard, 2018; Roediger & Butler, 2011; Rowland, 2014; Schwieren et al., 2017; Yang et al., 2021). Thus, retrieving information, actively answering test questions, solving test problems, and generating solutions to test questions is more beneficial than passively re-reading the same materials, re-studying, note-taking, or concept mapping (especially when feedback is provided or mistakes are corrected; e.g., Adesope et al., 2017; Agarwal et al., 2021; Batsell et al., 2017; Dunlosky et al., 2013; Karpicke & Blunt, 2011; Lechuga et al., 2015; Rummer et al., 2017; Yang et al., 2021). These long-term learning effects of tests were obtained for varying—complex, curricular, and difficult—topics (e.g., biology, engineering, history, language, mathematics, or psychology), in different online or face-to-face learning settings (e.g., schools, universities, laboratories, or at home/outside of class), and for students of different age groups (e.g., elementary school students, high school students, or university students; e.g., Adesope et al., 2017; Agarwal et al., 2021; Karpicke, 2017; Karpicke & Aue, 2015; McDaniel et al., 2011; Rawson, 2015; Roediger et al., 2011; Rowland, 2014; Yang et al., 2021). Benefits of test were also found when applying a wide range of test question formats (e.g., free recall tasks, multiple-choice questions, short answer questions, application-based questions, or transfer questions) and for varying types of learning materials presented during the initial learning opportunity (e.g., word pairs, vocabulary, factual information, conceptual information, longer textbook paragraphs, live lectures/lessons, or recorded e-lecturers/video-presentations; e.g., Adesope et al., 2017; Batsell et al., 2017; Feraco et al., 2020; Heitmann et

al., 2018; Jing et al., 2016; Khanna, 2015; McDaniel et al., 2013; Pan & Rickard, 2018; Roediger & Karpicke, 2006; Rowland, 2014; Yang et al., 2021). Notably, the application of tests was even effective when tests were administered with different modalities (e.g., orally, with paper-pencil, with computers, on online-websites, using clicker response systems, with mobile devices, or with games or game-based online applications; see e.g., Feraco et al., 2020; Iwamoto et al., 2017; Mavridis & Tsiatsos, 2017; McDaniel et al., 2013; A. I. Wang & Tahir, 2020; Yang et al., 2021).

Concerning the theoretical basics of tests and generation tasks—and of desirable difficulties in general—, it is often argued that their beneficial long-term learning effects arise because their higher difficulty triggers the accumulation of more (cognitive) effort and more (cognitive) resources: Expending more effort and resources to work on such difficulties and to overcome such challenging and demanding learning tasks in turn stimulates cognitive processes that then increase deeper (semantic, systematic, and cognitive) processing, encoding, and understanding of the to-be-learned information (e.g., Alter et al., 2007; R. A. Bjork, 1994; Bjork & Bjork, 1992, 2011; Craik & Tulving, 1975; Dunlosky et al., 2013; Pyc & Rawson, 2009; Roediger & Karpicke, 2006; Tyler et al., 1979). Tests and generation tasks are also assumed to lead to deeper and more effortful retrieval practice, to more elaboration, and to more analytic and elaborative reasoning/thinking (e.g., R. A. Bjork, 1994; Bjork & Bjork, 1992, 2011; Carpenter & DeLosh, 2006; Dunlosky et al., 2013; McCurdy et al., 2020; Rowland, 2014). They further anchor the learned information in long-term memory, connect the retrieved/generated information with already stored information, lead to more memory consolidation, and generally strengthen memory paths, traces, and associations (e.g., Bjork & Bjork, 1992, 2011; Carpenter, 2009; Gardiner & Hampton, 1985; Hirshman & Bjork, 1988; Karpicke et al., 2014; McCurdy et al., 2020; Roediger & Karpicke, 2006; for good overviews of different theories explaining the beneficial effects of tests and generation tasks see also: Karpicke, 2017; McCurdy et al., 2020). Moreover, the respective literature also highlighted

the importance of learners' successfulness while working on desirable difficulties: More specifically, tests and generation tasks were found to be more beneficial the more test questions learners could successfully answer, the more information they could successfully retrieve, and the more information they could successfully generate (see e.g., Abel & Hänze, 2019; Bjork & Bjork, 2019; S. Greving & Richter, 2018; Kaiser et al., 2018; Richland et al., 2005; Rowland, 2014; Sotola & Crede, 2021). It was accordingly also shown that higher initial test performances were crucial for later long-term learning effects of tests, insofar as that learners did not profit from tests when their initial test performance was low and they did not receive feedback (see e.g., Kang et al., 2007; McDaniel et al., 2007; Sotola & Crede, 2021; fortunately, failures or errors while retrieving or generating can still result in benefits, especially when mistakes are corrected or when feedback is given, see e.g., Bjork & Bjork, 2019; Kang et al., 2007; Kornell et al., 2009; Kornell & Vaughn, 2016; Potts & Shanks, 2014). Apart from learners' successfulness, it was also emphasized that the benefits of tests and generation tasks increase with higher (cognitive) effort and less support during retrieval and processing, with higher quality and depth of processing and encoding, and generally with higher difficulty of the tests or the generation tasks (e.g., Alter et al., 2007; Bertsch et al., 2007; Bjork & Bjork, 1992; Endres & Renkl, 2015; Karpicke, 2017; Karpicke & Roediger, 2007; Pyc & Rawson, 2009; Rowland, 2014; Tyler et al., 1979). It was accordingly also shown that difficult successful retrieval elicited more long-term learning benefits than easier successful retrieval and that more difficult test question formats and questions that increase the depth of retrieval were more beneficial than easier question formats and questions triggering only more shallow retrieval (e.g., S. Greving & Richter, 2018; Maass & Pavlik, 2016; Pyc & Rawson, 2009). Hence, to be beneficial, tests and generation tasks must be challenging, difficult, and effortful but should thereby still be solvable and not overwhelming (see also: Bjork & Bjork, 2019). Karpicke (2017) also noted that a balance between the

successfulness of learners' retrieval and the effort they must expend to retrieve the information must be given for tests to be beneficial.

However, especially these required increases in effort and difficulty often serve as the basis for lecturers' worries about the effectiveness of desirable difficulties, for lecturers' concerns regarding negative impacts of test, and for learners and lecturers previously described misconceptions concerning beneficial effects of difficult learning strategies (see e.g., Bjork & Bjork, 2019; Diemand-Yauman et al., 2011; Kirk-Johnson et al., 2019; Lipowsky et al., 2015; Yang et al., 2021). Thus, not all lecturers or learners might directly perceive desirable difficulties as positive, helpful, or worth the struggle and not all learners might in turn be motivated to try to overcome the posed difficulties or to extort more effort and more cognitive resources while learning. Besides, not all of them might be even able to increase their effort or resources and to successfully work on such difficult tasks. In line with these considerations, researchers previously assumed that less motivated or less able learners might rather give up when trying to solve difficult tasks instead of persisting, mustering more effort, or encoding the information more deeply (see e.g., Diemand-Yauman et al., 2011; McNamara et al., 1996). It hence appears valuable to investigate potential prerequisites or boundary conditions for the successful application of desirable difficulties. In the following, I will therefore briefly describe different cognitive-motivational learner characteristics that are linked to learners' attitudes regarding desirable difficulties and to the beneficial long-term learning effects elicited by these tasks. I will then turn to a more thorough exploration of learners' cognitive abilities and intelligence as prerequisites for the effectiveness of tests and generation tasks.

Prerequisites for Desirable Difficulties

Cognitive-Motivational Learner Characteristics

Regarding potential prerequisites for desirable difficulties, it generally seems to be important that learners appreciate hard work, difficult tasks, or challenges and that they believe that they will be able to reap their benefits. Otherwise, learners might have unfavourable views or unfavourable attitudes towards desirable difficulties, might not use them, might not give their best while learning with them, might not extort more effort, and might not even try to successfully retrieve, generate, or process the learned information. Accordingly, previous research showed that learners that are more appreciative of challenging learning tasks and cognitive engagement (thus, learners with higher levels of need for cognition) and learners that study with the intention to thoroughly understand and master the presented learning materials (thus, learners with higher levels of mastery goal orientation) had more positive attitudes towards desirable difficulties and reported to use them more often (e.g., Weissgerber et al., 2016, 2018). Similarly, learners that generally feel more challenged and less threatened in different test situations or by different test tasks (thus, learners with lower levels of trait test anxiety) held more positive attitudes towards tests, reported to use them more often (compared to repetition learning strategies), and had, in turn, better grades (e.g., Weissgerber & Reinhard, 2018). Following these previous findings, we hypothesized that learners with higher beliefs to be able to control their learning outcomes, with higher confidence in their success while working with desirable difficulties, and with generally less stress perceptions in different demanding (learning) situations should have more positive attitudes towards desirable difficulties (including more favourable views and more perceptions of usefulness) and should use them more often (Reinhardt et al., 2019; see Appendix A): As assumed, our online study (N = 504) yielded that higher internal locus of control and higher self-efficiency were linked to more positive attitudes towards different desirable difficulties and to higher self-reported usage. Our results further showed that participants' trait stress was negatively correlated to their attitudes towards desirable difficulties and to their self-reported application. When focusing on the specific types of

desirable difficulties separately instead of on an accumulated score, these negative correlations were especially distinct for tests. Similarly, Rivers (2021) later showed that learners' use of tests was dependent on their beliefs of success, insofar as that they reported to only use tests instead of re-reading tasks when they believed that they would be able to successfully retrieve the answers to the test questions. In contrast, they would choose re-reading when they believed that the retrieval of the correct answers would be too difficult and too challenging. Interestingly, it was recently shown that even at first glance farfetched learner characteristics that are linked to preferences for hard work and effort (like conservate attitudes) were in turn also related to positive attitudes towards desirable difficulties and to difficult learning in general (Mariss et al., in press). Hence, previous and own work implies that (cognitive-motivational) learner characteristics can act as prerequisites for learners' attitudes towards desirable difficulties and for their self-regulated application of those. In turn, such characteristics might also be linked to long-term learning outcomes elicited by desirable difficulties: For instance, learners' performance expectancies—that are linked to difficulty perceptions of learning tasks and that serve as amalgamations of subjective ratings, expectancies, and beliefs how well learners will be able to manage varying tasks (e.g., Dickhäuser & Reinhard, 2006; Eccles & Wigfield, 2002; Marshall & Brown, 2004)—might also influence learners' effort, motivation, and persistence while working with desirable difficulties as well as the thereby resulting benefits. Thus, to test these assumptions, a classroom study was conducted that focused on linkages between participants' performance expectancies and the long-term learning effects of generation tasks compared to reading already solved examples as an easier control task (N = 61; Reinhard et al., 2019; see Appendix B). The results of this study showed that after a delay of 3-month the assumed long-term learning effects of generation tasks only arose for participants with lower to average initial performance expectancies. Participants with higher initial performance expectancies only benefited immediately from generation tasks but not after the delay. A recent study from

Heitmann, Grund, et al., (2021) in contrast yielded that learners' hope of success moderated the effectiveness of tests, insofar as that tests were more beneficial for learners with higher hope of success and less beneficial for learners with lower hope of success.

Taken together, the just described findings imply that different cognitive-motivational learner characteristics can not only act as prerequisites for learners' perceptions of desirable difficulties but also for the benefits learners are able to obtain through the application of such difficulties. Some studies thereby highlighted the importance of higher cognitive motivation to engage in difficult and challenging learning, of higher hopes of succeeding while working on difficult tasks, and of higher beliefs to be able to control and master their learning outcomes and the respective difficulties (see e.g., Heitmann, Grund, et al., 2021; Reinhardt et al., 2019; Weissgerber et al., 2016, 2018). In contrast, other studies instead highlighted the importance of lower cognitive motivation or of lower performance expectancies (see e.g., Reinhard et al., 2019; Schindler et al., 2019). A recent study even found no linkages among learners' need for cognition, their grit, and the effectiveness of tests for later learning outcomes (e.g., Bertilsson et al., 2021). These ambiguous results thus indicate that further research focusing on linkages among cognitive-motivational learner characteristics, attitudes towards desirable difficulties, and long-term learning effects of desirable difficulties is still needed. Moreover, these contrary findings also imply that future research should not only focus on learners' beliefs and expectancies to be able to manage difficult tasks but also—and even more thoroughly—on learners' actual abilities to successfully work on desirable difficulties. Bjork and Bjork (2019, p. 166) accordingly stated that: "Desirable difficulties are desirable because responding to them (successfully) engages processes that support learning, comprehension, and remembering. They become undesirable difficulties if the learner is not equipped to respond to them successfully." Hence, I will now discuss which (and to what extent) cognitive abilities and intelligence serve as prerequisites that must be given for

learners to be adequately equipped to successfully respond to desirable difficulties and to reap their benefits.

Cognitive Abilities and Intelligence

Given the required difficulty of tests and generation tasks, responding successfully to desirable difficulties—indicated by successful initial performance, successful retrieval, and successful generation—is probably not automatically possible for all learners, particularly when learning complex materials. Considering that higher intelligence has often been shown to be strongly linked to better long-term memory, higher retrieval success, higher initial test performance, higher long-term learning outcomes, higher academic achievement, and higher complex problem solving (e.g., Fellman et al., 2020; Fergusson et al., 2005; Kuncel et al., 2004; Roth et al., 2015; Stadler et al., 2015; Stern, 2015; Strenze, 2007; Unsworth, 2019; T. Wang et al., 2017), such successful responding should, however, be likely for learners with higher intelligence. It has additionally been argued that learners can only reap the benefits of desirable difficulties if they can even muster the required increased effort, the extended thought, and the more elaborated, analytical, or effortful processing (e.g., Alter et al., 2013; Oppenheimer & Alter, 2014; see also the *aptitude-treatment-interaction* or the *expertise-reversal effect*: e.g., Kalyuga et al., 2003; McDaniel & Butler, 2011; Snow, 1989). Higher intelligence should increase this possibility. Researchers hence assumed that the beneficial effects of desirable difficulties arise mostly for those learners that can extort the needed increased effort and that can successfully overcome the heightened difficulty, but additionally also only for those learners that are not cognitively overwhelmed and do not have to deplete all their cognitive resources while doing so (e.g., Alter et al., 2013; Bjork & Bjork, 2011; Kalyuga et al., 2001; Kornell et al., 2011; McDaniel et al., 2002; Oppenheimer & Alter, 2014; Richland et al., 2005; Rowland, 2014). McDaniel and colleagues (2002) thereby specified that not only learners' ability to successfully cope with desirable difficulties is relevant for their

beneficial effects, but also the amount of cognitive resources learners have left after working on them: The authors found that both less able as well as more able readers could successfully solve the posed generation tasks, but that only more able readers actually benefitted from these tasks in the long-run—potentially because these learners did not have to use up most or all of their processing capacities to correctly solve the tasks but still had cognitive resources left to further process and deeper encode the correctly generated information. Because higher intelligence is strongly linked to higher cognitive resources and to more successful and effective (cognitive) information processing (e.g., Bornstein et al., 2013; Gottfredson, 1997; Oberauer et al., 2005; Stern, 2015, 2017; Unsworth et al., 2014; T. Wang et al., 2017), the same should apply to learners with higher intelligence. Hence, learners with higher intelligence are assumed to benefit from desirable difficulties more strongly because they should be able to work on them more successfully than learners with lower intelligence and should additionally still have enough resources left to process the information more deeply—even after working on such difficult and cognitive capacities reducing tasks. Moreover, previous work generally noted the relevance of higher cognitive abilities, higher cognitive resources, higher knowledge, and higher achievement for obtaining the benefits of desirable difficulties: More specifically, learners with higher working memory capacities, higher prior knowledge, more experience, more expertise, higher reading abilities, higher spelling skills, and those that were generally high achieving were shown to benefit (especially) from desirable difficulties (e.g., Carpenter et al., 2016; Eskenazi & Nix, 2021; Kalyuga et al., 2001; Lehmann et al., 2016; McDaniel et al., 2002; McNamara et al., 1996). Notably, (higher) intelligence was also often found to be strongly related to these cognitive variables (see e.g., Bornstein et al., 2013; Fergusson et al., 2005; Gottfredson, 1997; Oberauer et al., 2005; Stern, 2015, 2017; Sternberg, 1997; Unsworth, 2010; T. Wang et al., 2017).

These findings and argumentations indicate that sufficient or higher intelligence might serve as a prerequisite or boundary condition for the long-term learning effects of tests and

generation tasks—especially with regard to complex and curricular materials. Nonetheless, although further research investigating these assumptions is needed and requested (see Dunlosky et al., 2013), only a small amount of work has until now been conducted to investigate these research issues. Kaiser et al. (2018) for instance focused on linkages between school students' intelligence (indicated by figural inductive reasoning), generation tasks, previous knowledge, and long-term learning using complex materials in a realistic inquiry-based learning setting: They found that higher intelligence was linked to higher previous knowledge, which was in turn linked to higher long-term learning outcomes after 1-week. Another study yielded that college students with higher general fluid intelligence benefitted more form tests (compared to re-reading) after a delay of 2-days when the learning materials (normed Swahili-English vocabulary) were difficult but not when they were easy (Minear et al., 2018). Difficult information probably made them increase their effort in order to answer the test questions successfully, which more intelligent learners were still able to manage, whereas the easy information were probably too easy for these learners and did thus not trigger the needed increase in effort that elicits the benefits of tests. In contrast, less intelligent learners only befitted from tests when the information were easy, probably because for them easy information already triggered enough additional effort that they were still able to successfully overcome, whereas difficult information were probably too difficult and only overwhelming for these learners (Minear et al., 2018). However, previous work also resulted in contrary findings: For instance, Brewer and Unsworth (2012) showed that when learning word pairs university students with lower general fluid intelligence benefitted more from tests after a delay of 1-day than learners with higher intelligence (although learners with higher intelligence generally performed better and had higher long-term learning outcomes than learners with lower intelligence). Moreover, elementary school students' processing speed (which serves as one aspect of cognitive abilities and fluid intelligence) did not moderate the effectiveness of retrieving vs. re-reading word lists after a short delay (Karpicke et al., 2016).

Robey (2017) also found no moderating effect of university students' general fluid intelligence on their learning outcomes after a delay of 30-minutes following either tests retrieving initially learned word pairs or re-studying of these words. These contrary results might be explained by the different levels of complexity or difficulty of the used learning materials and by the different delays measuring later learning outcomes: It is possible that higher intelligence would only be a prerequisite for the beneficial effects of tests when using more difficult and complex learning materials and after longer delays.

However, due to these varying findings, the generally few conducted studies, and the importance of this research issue for later applications (e.g., regarding potential boundary conditions describing for whom tests or generation task should be applied and for whom not), more empirical research is needed. This applies especially to research using longer delays and difficult, complex, and curricular materials as implemented in school classes or university courses: Thus, we conducted two laboratory studies testing the assumption that (higher) intelligence moderates the effectiveness of generation tasks and tests (Wenzel & Reinhard, 2019; see Appendix C). Both studies included university students as participants (Study 1: N = 149; Study 2: N = 176), measured intelligence using a valid and detailed intelligence test assessing overall intelligence and reasoning (*Intelligence Structure Test*, I-S-T 2000 R: Liepmann et al., 2007), and assessed participants' prior knowledge concerning the respective learning materials before the actual learning phases started (due to the often reported relevance of prior knowledge, see e.g., Bjork & Bjork, 2011; Kalyuga et al., 2001; McNamara et al., 1996; Stern, 2015). Long-term learning outcomes were assessed after 2-weeks (Study 1) and after 1-week (Study 2). In the first study, participants read basic information on linear regressions and then either learned through working on generation tasks (e.g., actively filling blanks, sketching a function into a graph, or generating solutions to mathematical questions) or through reading the already solved tasks. The results showed that participants' intelligence was positively correlated with their prior knowledge and with their long-term learning (even

beyond and under control of prior knowledge). There was, however, no beneficial effect of the generation tasks on participants' later learning outcomes and there was also no moderating effect of intelligence on the effectiveness of the learning tasks. Notably, an applied manipulation check showed that participants did not perceive the generation tasks as more difficult than the reading control tasks, indicating that our manipulation of the learning situation was not successful to begin with. Hence, the generation tasks were apparently not challenging or demanding enough to elicit the required increased effort and the deeper cognitive processing that normally trigger the benefits of generation tasks. Although this contradicted most previous work (e.g., Bertsch et al., 2007; McCurdy et al., 2020), it fitted some research that did also not continuously result in positive effects of generation (see e.g., de Winstanley & Bjork, 2004; de Winstanley et al., 1996; Karpicke & Zaromb, 2010; Metcalfe & Kornell, 2007). Some researchers even argued that the application of tests is probably more robust than the application of generation tasks (e.g., Karpicke & Zaromb, 2010). Therefore, our second study (Wenzel & Reinhard, 2019; see Appendix C) tested our hypotheses using short tests instead of generation tasks. As an initial study opportunity participants once read a university textbook chapter regarding biopsychology and the lateralization of the brain. Thereafter, they were given 10-minutes to either re-read the text as often as they could or to answer multiple test questions (they later received the correct answers as short feedback). A conducted manipulation check showed that participants perceived the test as more difficult than the re-reading control task, thus showing that the manipulation of our learning condition was successful. In line with previous work, the results of this second study yielded positive correlations among participants' intelligence, their prior knowledge, and their learning outcomes. We also found that tests increased participants' later learning outcomes compared to the easier re-reading control task, thus supporting the desirability of tests. This beneficial effect was, in turn, moderated by participants' intelligence, insofar as that tests were not beneficial for participants with relatively low

intelligence but increased long-term learning of participants with average intelligence. Participants with relatively high intelligence benefitted even more from taking a test. This effect remained robust even when the positive effects of prior knowledge were controlled for. This indicates that the found interaction-effect was not simply due to the benefits of higher prior knowledge but that higher intelligence (and its, among others, higher cognitive resources, better and faster information processing, deeper and more successful retrieval, and more analytical or abstract thinking) is a valuable prerequisite for the effectiveness of tests. Taken together, our results again show the robust beneficial effects of short tests—even when including different test question formats (like short-answer questions and multiple-choice questions), when assessing varying depths of knowledge (like factual knowledge and transfer knowledge), and when using realistic and complex learning materials (see also: Adesope et al., 2017; Agarwal et al., 2021; Pan & Rickard, 2018; Rowland, 2014; Yang et al., 2021). In addition, our results also supported the theoretical and empirical assumptions that tests might not be beneficial for every learner but that intelligence might act as a boundary condition for the effectiveness of tests: Although the effect was small and although replications and further work (e.g., focusing even more closely on the specific aspects of intelligence) are still valuable, our results nonetheless yielded that at least average intelligence needs to be given for learners to be able to reap the benefits of tests. Fortunately, the less intelligent participants did at least not suffer under the application of tests (there was no *poor-get-poorer effect*, see e.g., Stanovich, 1986): Even though less intelligent participants using tests did not outperform similarly less intelligent participants using re-reading tasks, their later learning outcomes were at least not worse than the learning outcomes of similar participants using the easier control task.

Focusing purely on later learning outcomes, these results would hence imply that tests should be applied in actual learning context because they would be beneficial for most learners and would not change or decrease the learning outcomes that the remaining learners

would have achieved anyways using easier tasks. However, the application of tests and their desirability can—and should—also be considered and evaluated beyond long-term learning outcomes. Independent of the gains learners can reap by (successfully) working on tests, they all thereby have to undergo a demanding and challenging learning task, have to work harder compared to the re-reading task, and have to extort more effort to try to answer the test questions—and even the learners that profit thereof cannot immediately observe the reaped benefits but might only realize that the tests had been beneficial after a (longer) delay. Considering that most learners often mistakenly believe that effective learning feels easy, that a need to increase the effort to solve a task equals inadequate cognitive abilities, that immediate performance corresponds to later learning outcomes, and that tests are only effective for self-evaluations or assessments (see e.g., Bjork & Bjork, 2019; R. A. Bjork et al., 2013; Kornell et al., 2011; Miele et al., 2011; Muenks et al., 2016), such difficult learning situations might thus be perceived as unpleasant or not worth the struggle. Hence, the application of tests and the ensuing difficult and demanding learning situation might result in negative perceptions, experiences, or evaluations immediately during learning. It is thus extremely valuable to explore if tests lead to such acute negative side-effects compared to easier or more fluent leaning situations.

Acute Negative Side-Effects of Tests as Desirable Difficulties

In line with these considerations, previous research has often found that more difficult tasks generally increase perceptions of threat or anxiety, that experiencing difficulties or giving incorrect answers feeds negatively into self-perceptions, and that (subjectively) performing poorly results in higher stress perceptions and elicits more test anxiety (e.g., O'Neil et al., 1969; Ott, 2017; Sarason & Sarason, 1990; Schunk & Gaa, 1981). More difficult tasks and learning tasks that require more effort, more time, or more workload were also shown to be perceived as more stress-inducing compared to easier tasks (e.g., Kausar, 2010).

Theoretically, such acute stress perceptions normally arise in situations—or when working on tasks—that are perceived as threatening and overwhelming instead of challenging and when individuals think that they do not possess enough cognitive resources or abilities to manage the posed demands (see e.g., the *transactional theory of stress*: Lazarus, 1990; Lazarus & Folkman, 1984, 1987). This should also apply to tests as desirable difficulties, which are even designed to be difficult, demanding, and effortful and might thereby be perceived as threatening and stressful tasks, as too demanding, overwhelming, and consuming too many resources. Because desirable difficulties also reduce leaners' *illusions of competence* and their overconfidence (see e.g., Alter et al., 2007; R. A. Bjork, 1999; Koriat, & Bjork, 2006), learners get an adequate but potentially unsatisfactory view of their learning progress, which may further result in perceptions of imbalances between the difficult tasks and learners' resources or capabilities. In turn, such perceptions often elicit stress and anxiety (e.g., Bystritsky & Kronemyer, 2014; Endler, 1997; Epel et al., 2018; Hobfoll, 1989; McGrath, 1970). Leaners were also shown to perceive retrieval failure (which is likely to occur when working on tests) as a form of negative feedback, to rate tasks including retrieval as less enjoyable than tasks including re-reading, and to report that re-reading (but not working on tests) feels good (see e.g., Biwer et al., 2020; Clark & Svinicki, 2015; Rivers, 2021). Previous research fittingly showed that tests can lead to cognitive overload, to higher general anxiety, to increased test anxiety, and to more stress (e.g., Hinze & Rapp, 2014; Khanna, 2015; van Gog & Sweller, 2015). More specifically, Hinze and Rapp (2014) conducted a laboratory study using realistic science texts as study materials and found that short low-stakes learning tests led to more immediate feelings of pressure than re-reading tasks. High-stakes learning tests (operationalized through instructions stating that monetary rewards for the learner and a fictive partner were dependent of learners' later learning outcomes) further led to more state anxiety compared to the low-stakes tests without these instructions and compared to the easier re-reading tasks. These effects arose for all learners regardless of their trait test anxiety (Hinze

& Rapp, 2014). Thus, these findings indicate that negative side-effects like increased feelings of pressure, anxiety, or acute stress perceptions are not restricted to high-stakes situations like examinations but can also—unfortunately for the respective learners—evolve in varying low-stakes learning situations when using normally beneficial and helpful learning tests. Notably, contrary results showing that tests had no effects on test anxiety or that they reduced test anxiety concerning later examinations were also found (see e.g., Agarwal et al., 2014; Nyroos et al., 2016; Szpunar et al., 2013; Yang et al., 2020). Although these varying results might be caused by content-related or methodological differences (e.g., concerning the focus on later examinations instead of on tests themselves, concerning the use of delayed or even retrospective assessments of stress and anxiety instead of immediate assessments while using tests, or concerning the application of abstract word lists as learning materials instead of more naturalistic or curricular learning materials), further research is therefore extremely valuable.

Given these varying results and given that learners' emotions and perceptions during varying learning situations are seldom the main focus of studies (see e.g., Edwards & Templeton, 2005; Goetz et al., 2007; Rauthmann, 2012), it is relevant to further focus on learners' experiences while learning with tests before these are recommended to be applied in schools or universities. Such further work thereby seems especially important given that higher stress perceptions—that include nervousness, tension, anxiety, feelings of pressure, feelings of overwhelm, worry, intrusive and disturbing thoughts, lack of confidence, and subjective distress (e.g., Epel et al., 2018)—are extremely unpleasant and are normally avoided during learning. However, although required, there is not much research investigating if (and to what extent) tests negatively affect learners' immediate stress perceptions during learning, their acute negative evaluations of the learning situation, and their anxiety experiences while learning. We therefore conducted two studies to investigate negative side-effects potentially caused by tests (Study 1: $N = 405$; Study 2: $N = 102$; Wenzel & Reinhard, 2021a; see Appendix D). We firstly conducted an online study that measured different learner

characteristics (trait stress and trait anxiety) and used hypothetical scenarios that participants were instructed to read and imagine. The short scenarios described learning situations spanning a whole hypothetical semester (ending with an important examination) in which a lecturer either ended every session with a re-reading task, a learning test with private results (that were sent individually per mail to every student), or a learning test with more public results (that were later sent simultaneously to all students depicting their matriculation numbers and test results). After reading, participants reported their negative evaluations of the imagined learning situation (including items assessing feelings of unfairness, anger, uncertainty, annoyance, and pressure, as well as ratings of difficulty, injustice, and strenuousness) and the acute stress they would perceive in such a learning situation. The results yielded that both scenarios including tests were evaluated as more negative than the scenario including re-reading and that the learning scenario including public test results was evaluated the most negative. Notably, lower trait stress was able to buffer participants' negative evaluations caused by the learning scenario including tests with private results but could not buffer the negative evaluations caused by the learning scenario including tests with public results. Hence, this scenario was continuously evaluated more negatively than the other scenarios regardless of participants' levels of trait stress. Concerning their state stress, we unexpectedly found no significant effects of the learning scenario conditions, indicating that the scenarios including tests were not perceived as more stressful than the re-reading control scenario. However, it is possible that the scenarios were not detailed, long, or realistic enough to elicit actual experiences of stress. This assumption indicated that further studies should use longer and more realistic scenarios or should be directly conduced in laboratory or classrooms settings. Thus, our second study (Wenzel & Reinhard, 2021a; see Appendix D) was conducted in a laboratory and set up like a typical university seminar room. We again assessed learner characteristics (traits stress, trait anxiety, and task-specific self-concept) to investigate if the expected negative side-effects would arise for all participants. After a short

initial study opportunity on linear regressions, participants then either worked on a short test (they thereby also had to answer two of the test questions out loud in front of the other participants) or simply read the answers of the already solved tasks. Directly afterwards, participants' acute stress perceptions, their feelings of anxiety during learning, and their negative evaluations of the learning situation were assessed. In line with the findings of our first study and in line with previous work (e.g., Hinze & Rapp, 2014), our results implied that the learning situation including the test led to higher perceptions of stress and anxiety and to more negative evaluations of the situation compared to the learning situation including the easier reading control task. These negative side-effects were independent of participants' trait stress or their trait anxiety. Although the found effects were not extremely strong, these results nonetheless indicate that learning situations including tests can result in relatively unpleasant experiences for the learners working with them. It seems thus important to consider that although tests are often desirable for later learning outcomes, they can also be undesirable due to negative side-effects and unpleasant immediate experiences they cause. Notably, these negative side-effects arose even though the applied tests were only short and low-stakes, even though learners' initial test performance had no consequences for their studies or everyday lives, even though participants were only instructed to do their best while learning without being given further incentives, and even though the learning situation only took place in a laboratory setting. In turn, tests applied in actual schools or universities (including the presence of peers, realistic and complex learning materials that are part of the curriculum, and actually relevant later learning outcomes or grades) might cause even stronger negative side-effects. Future research testing this assumption is thus valuable. Such future research could, among others, also generally try to replicate our findings (e.g., using different types of test question formats, varying test application modes, different control conditions, or further types of negative side-effects) and try to explore these negative side-

effects in more detail (e.g., concerning how long-lasting they are, if they remain robust even after repeated applications, or how the effects can be reduced).

An additional concern that needs to be considered is that applications of tests might not only lead to the just described negative side-effects immediately during learning, but also—directly or indirectly via those negative side-effects—to further undesirable consequences after a delay. Following, I thus want to focus on two potential instances of such further negative consequences: increased academic cheating and impaired effectiveness of tests resulting in reduced later learning outcomes.

Further Negative Consequences of Tests as Desirable Difficulties

Academic Cheating

In line with this consideration, previous studies yielded that learners' perceptions of courses or assessments as (too) difficult increased academic misconduct and that the difficulty of the course also served as a reason to justify or rationalize cheating (e.g., Brimble & Stevenson-Clarke, 2005; Freiburger et al., 2017; Haines et al., 1986; Steininger et al., 1964). Higher extorted effort and higher workload were also linked to feelings of entitlement, moral justifications, and cheating (e.g., Hoffman & Spitzer, 1985; McCabe, 1992; Whitley, 1998). Hence, tests as desirable difficulties might directly lead to higher intentions to cheat and to more justifications for academic cheating. Additionally, tests might also indirectly lead to more academic cheating via the negative side-effects they cause: More specifically, increased stress perceptions, higher feelings of anxiety, and more negative situation evaluations (including, among others, negative affective states or feelings of unfairness) were often found to be linked to academic cheating, to intentions to cheat, or to justifications for cheating (see e.g., Agnew, 1992; Houser et al., 2012; Olafson et al., 2013; Wowra, 2007). Stress perceptions, anxiety (including test anxiety, social anxiety, and general anxiety), and feelings of pressure (including general feelings of pressure, pressure for grades, and parental pressure)

were also shown to be related to cheating or were reported as reasons and incentives for such dishonesty (e.g., Brimble & Stevenson-Clarke, 2005; Davis et al., 1992; Rost & Wild, 1994; Schab, 1991; Whitley, 1998; Wowra, 2007). The more a test situation was further experienced as pressuring and uncomfortable, the more unfair the testing tool was perceived (Leiner et al., 2018)—and unfairness (of teachers or of teaching practices) was, in turn, also correlated with (academic) cheating and justifications for cheating (see e.g., Brimble & Stevenson-Clarke, 2005; Finn & Frone, 2004; Freiburger et al., 2017; McCabe, 1992; Olafson et al., 2013; Whitley, 1998). These findings indicate that the applications of tests might directly as well as indirectly trigger more cheating in the academic context and serve as justifications for such cheating.

Given that cheating is undesirable behaviour that should neither occur nor be normalized in academic settings and that further negatively impacts fair grading and performance successions in class (see e.g., Carrell et al., 2008; Fida et al., 2018; Gino et al., 2009; McCabe et al., 2001; Paternoster et al., 2013), it is extremely important to investigate if cheating can be triggered by applying tests in schools or universities. Because there were to our knowledge no previous studies focusing on linkages among tests, negative side-effects, and cheating, we conducted an online study that investigated if (and to what extent) academic cheating might be directly or indirectly caused by tests (N = 405; Wenzel & Reinhard, 2020; see Appendix E). Participants thereby imagined one of three scenarios that described learning situations taking place throughout a whole semester and culminating in an examination. Two of these learning scenarios included tests, whereas the third learning scenario described easier re-reading. Participants then reported their negative evaluations of the respective scenario (including items assessing feelings of unfairness, anger, uncertainty, annoyance, and pressure, as well as ratings of difficulty, injustice, and strenuousness) and the acute stress they would perceive in such a situation. Subsequently, participants read another short scenario describing the examination at the end of the hypothetical semester and answered different items

concerning likelihoods of cheating as well as their justifications for such cheating behaviour: Participants were thereby instructed to rate how likely it would be that they would (spontaneously or with preparation) cheat in this hypothetical examination, how likely it would be that their peers would (spontaneously or with preparation) cheat during the examination, and how justified the respective cheating would be. Mediation analyses found neither direct effects of the learning scenarios on participants' immediate stress perceptions, nor on their likelihoods for cheating, nor on their justifications for such hypothetical cheating. However, there were significant indirect effects: Tests indirectly led to higher likelihoods that participants would cheat in an examination situation as well as to higher justifications therefore by increasing their negative evaluations of the learning situation. More specifically, the conducted mediation analyses showed that the learning scenarios predicted participants' negative evaluations of the learning situation, insofar as that the scenarios including tests were evaluated as more negative than the scenario including re-reading. In turn, these increased negative evaluations were then linked to higher likelihoods of participants' own cheating and to their general justifications for cheating. Hence, in line with the presented empirical and theoretical assumptions, these findings indicate that tests as normally beneficial learning strategies not only cause negative side-effects indicated by negative situation evaluations but can thereby also indirectly trigger further negative consequences like higher likelihoods of academic cheating and higher justifications for such cheating. Although these indirect effects can only be classified as small to medium, they are still extremely intriguing because they arose even though we solely used short hypothetical learning scenarios. Genuine leaning situations including real learning materials and actual incentives to do well in later examinations should thus result in even stronger effects. This is why I argue that it is important to highlight these novel results that uniquely bridge the until now unconnected literature concerning tests, their negative side-effect, and academic cheating. Given the lack of further empirical data, future research trying to replicate or broadening our findings is

definitely needed. This applies especially to studies investigating these research issues in applied learning settings using realistic materials and assessing observable instead of hypothetical cheating behaviour. Future work should also test causal effects in addition to these correlative findings as well as potential influences of learner characteristics (e.g., attitudes towards dishonesty). Focusing on ways to reduce these indirect effects of tests on cheating (e.g., by reducing learners' possibilities to cheat in later examinations or—more ideally—by stopping the emergence of negative side-effects earlier on) is also valuable.

These negative consequences indirectly caused by tests are extremely undesirable, especially given that cheating can, among others, further lead to unfair advantages for the cheating students, to wrong grading, to seemingly better learning outcomes, and to reduced validities of the examinations scores. These findings also appear to be particularly paradox considering that the application of a normally desirable learning strategy is linked to the later occurrence of (socially) undesirable behaviour. However, increased cheating might not be the only further negative consequence caused by tests: The by tests caused negative side-effects—especially the increased stress and anxiety perceptions—might also disrupt academic achievement, might impair long-term learning outcomes, and might even reduce the effectiveness of tests.

Impaired Effectiveness of Tests and Reduced Later Learning Outcomes

Accordingly, previous work yielded that (dispositional and situational) stress and anxiety were often related to lower motivation to learn, to more errors, to a lack of concentration, to disruptions in attention, to higher cognitive load, and to reduced effort and persistence while learning (e.g., Chen & Chang, 2009; Kurebayashi et al., 2012; LePine et al., 2004). Stress and anxiety were also linked to cognitive deficits and were repeatedly shown to be negatively correlated with cognitive information processing, effectiveness of retrieval practice, learning, and (test) performance—especially as the tasks and materials become more

complex, more cognitive demanding, and more difficult (e.g., Ashcraft & Krause, 2007; Cassady, 2004a, 2004b; Chen & Chang, 2009; Hembree, 1988; Khan et al., 2013; I. G. Sarason, 1984; Seipp, 1991; Sotardi et al., 2020; Struthers et al., 2000). Although contrary results indicating positive correlations among stress, anxiety, and learning outcomes also exist (see e.g., Keeley et al., 2008; LePine et al., 2004; Sung et al., 2016), stress and anxiety are mostly assumed to have detrimental effects on learning outcomes (see, among others, the *cognitive inference model*, *distraction theories*, the *processing efficiency theory*, and the *retrieval disruption hypothesis*: e.g., Ashcraft & Krause, 2007; Eysenck & Calvo, 1992; Eysenck et al., 2007; I. G. Sarason, 1984; Tobias, 1984; Wine, 1971). Notably, stress and anxiety were even shown to negatively impact and reduce the long-term learning effects of tests: For instance, Mok and Chan (2016) found that tests were only more beneficial than re-reading and summarizing tasks for learners with lower test anxious dispositions but not for learners with higher test anxious dispositions (for contrary findings see: Clark et al., 2018). Focussing on learners' situational perceptions, similar results were previously obtained by Hinze and Rapp (2014): Compared to re-reading tasks and low-stakes tests, high-stakes tests (operationalized through instructions stating that monetary rewards of learners and their fictive partners were dependent of higher later learning outcomes) increased learners' immediate anxiety perceptions and their acute feelings of pressure, which in turn disrupted the long-term learning benefits of these tests. Such high-stakes tests resulted in lower later learning outcomes than the low-stakes tests and only learners using low-stakes tests outperformed the re-reading condition in the long-run (Hinze & Rapp, 2014). These results indicate that acute perceptions of anxiety, pressure, or stress caused by tests might in turn mediate the beneficial effects of these tests, insofar as that higher scores might diminish (or even completely erase) the beneficial effects of tests on long-term learning. Because this would—in addition to creating unpleasant and undesirable learning experiences—completely

contradict the intention of applying tests, I considered it necessary to further focus on these assumptions.

It thereby appears to be further important to also focus on learners' cognitive abilities while investigating these research issues because higher intelligence might not only be an important prerequisite for the effectiveness of tests but might also buffer both the negative side-effects caused by tests as well as the thereby further triggered negative consequences. These considerations are, for instance, based on previous research indicating that negative side-effects caused by tests (like higher immediate stress and anxiety perceptions) are less distinct for more intelligent learners compared to less intelligent learners: More specifically, previous work showed that higher cognitive or intellectual abilities were often linked to lower subjective difficulty ratings of varying tasks, to lower acute stress perceptions, to lower math anxiety, and generally to lower state anxiety (e.g., Abín et al., 2020; Efklides et al., 1997; Goetz et al., 2007; LePine et al., 2004). Learners that were extremely high-achieving in mathematics also reported less math anxiety (García et al., 2016) and learners with higher abstract reasoning abilities reported more enjoyment compared to anger or anxiety during a mathematics achievement test (Goetz et al., 2007). Fittingly, learners that performed better during a learning test—hence, learners that were able to successfully retrieve more information—also enjoyed the test more (Clark & Svinicki, 2015). Moreover, previous research additionally implied that even if these negative side-effects would nonetheless arise, the thereby often triggered negative consequences (like reduced learning outcomes or disrupted effectiveness of tests) are also less distinct for more intelligent learners compared to less intelligent learners: It was, more specifically, argued that higher domain-specific abilities or extra processing resources can compensate the detrimental learning effects caused by stress and anxiety (e.g., Eysenck & Calvo, 1992; Eysenck et al., 2007; Naveh-Benjamin, 1991; Tobias, 1984; for contrary results see: e.g., Sung et al., 2016). Tse and Pu (2012) similarly showed that less successful retrieval practice caused by higher test anxiety could be

compensated by higher working memory capacities—thus, anxiety only had detrimental effects for learners with lower working memory capacities (see also: Ashcraft & Krause, 2007; Johnson & Gronlund, 2009; Owens et al., 2014; however, contrary results also exist: see e.g., Beilock, 2008; Yang et al., 2020). Reeve et al. (2014) also found that cognitive abilities buffered negative consequences of distraction: Thus, distraction negatively impacted performance of lower ability learners but did not decrease performance of higher ability learners. Fluid intelligence was also shown to moderate the negative effect of state anxiety on working memory functioning: State anxiety negatively affected working memory functioning for learners with intelligence below median, but this negative impact of state anxiety was shown to diminish with higher intelligence (Chuderski, 2014).

Jointly, these previous findings highlight the importance of combining the different research issues presented in this dissertation: It thus seems valuable to simultaneously focus on linkages among intelligence as a prerequisite of tests, on stress perceptions as a negative side-effect caused by tests, and on a reduced effectiveness of the applied tests as a negative consequence further caused by these side-effects. Because there is almost no research that simultaneously focuses on these issues—while also using realistic and curricular learning materials and when focusing on delayed instead of immediate learning outcomes—further work is still needed. Hence, we conducted a laboratory study to test these assumptions (unfortunately, our data collection had to be terminated prematurely due to the outbreak of the COVID-19 pandemic, which resulted in a reduced sample size: $N = 89$; Wenzel & Reinhard, 2021b; see Appendix F): We hypothesized tests to increase learners' long-term learning outcomes but also their acute stress perceptions, which should in turn reduce the effectiveness of tests on later learning outcomes. These three effects should additionally be moderated by intelligence: 1) tests should be more beneficial for more intelligent learners compared to less intelligent learners, 2) more intelligent learners should perceive less stress during the learning situation including a test compared to less intelligent learners, and 3) long-term learning

outcomes of more intelligent learners should be less impaired by stress than the long-term learning outcomes of less intelligent learners. Participants therefor underwent a short intelligence screening (measuring a general cognitive ability indicated by speeded reasoning as a conglomerate of reasoning, abstract thinking, and processing speed; *mini-q*: Baudson & Preckel, 2015) and then read a university textbook chapter on the brains' lateralization as an initial study opportunity (positively, this realistic learning material was part of the curriculum of a majority of the sample). They then either re-read the textbook chapter as an easier control task or took a short low-stakes test (including both short answer and multiple-choice questions) with later feedback. Directly thereafter participants' acute stress perceptions during the respective learning task were measured. Their long-term learning was later assessed after a delay of 1-week. In accordance with previous work (e.g., Adesope et al., 2017; Yang et al., 2021), our results again showed that taking tests increased participants' later long-term learning compared to re-reading. However, tests also again increased participants' immediate stress perceptions, thus leading to a more unpleasant learning situation (see e.g., Hinze & Rapp; Wenzel & Reinhard, 2021a). These increased stress perceptions were then, in turn, linked to reduced long-term learning outcomes. A mediation analysis thereby yielded that the by tests caused stress perceptions suppressed the beneficial effects of tests, insofar as that the beneficial effects were lower the more stress was perceived. Although these stress perceptions were not extremely strong and although tests were still beneficial in the long run, they could have been even more beneficial if they would have caused no or only average stress perceptions. These findings and these resulting considerations are especially irritating given that we had already intended—apparently without success—to conduct the applied test and the whole learning situation as low-stakes and stress-less as possible: Among others, participants knew that they would not have to give their answers out loud, that they would learn on their own without interacting with the other participants, and that their performances would have no consequences for their daily lives and would remain completely anonymous.

In turn, differently implemented tests or tests used in applied setting (including difficult materials, higher stakes, or consequences due to learners' later outcomes) might lead to higher stress perceptions and thus also to higher negative consequences concerning the effectiveness of the used test. Moreover, our study (Wenzel & Reinhard, 2021b; see Appendix F) further showed that participants' intelligence was positively correlated to their initial test performance and their later long-term learning. Intelligence was also negatively correlated to participants' stress perceptions but did not moderate any of the other three effects (hence, neither beneficial effects of tests on long-term learning, nor the negative side-effects of tests on stress perceptions, nor the detrimental effects of stress on long-term learning). Thus, although higher intelligence was generally advantageous, participants benefitted equally from tests and also suffered equally under the negative side-effects of tests and under the negative consequences of the increased stress perceptions. However, these findings must be interpreted with caution given that our small sample size should have been sufficient for testing the main and mediation effects but was probably too small for reliably testing the assumed interaction effects (see Blake & Gangestad, 2020). Hence, replications focusing on these supposed moderating effects of intelligence are urgently needed. Apart from that, our work generally emphasized the importance of trying to design tests so that they increase long-term learning outcomes without also resulting in higher stress perceptions (and thereby, in turn, without indirectly leading to a suppressed effectiveness of the respective tests). Hence, although this paper is not without limitations and can only be seen as a starting point or a first step, it nonetheless identifies valuable research issues. It also highlights the importance of a more thorough and encompassing look at the desirability of desirable difficulties and indicates—together with my other papers—that varying dimensions of desirability should simultaneously be considered when evaluating tests (or other desirable difficulties).

Discussion

In accordance with this consideration and with the title of my dissertation, the present work was conducted to take a closer look at the desirability of desirable difficulties in addition to their long-term learning effects. I therefore focused mostly on the application of tests as one of the most robust, helpful, and easily applicable type of desirable difficulties. In addition to the beneficial effects of tests, I also took intelligence as a cognitive prerequisite for their effectiveness, by tests triggered negative side-effects, and negative consequences caused indirectly by tests into account. I argue that these different factors—and their respective linkages—should be considered together when contemplating the general desirability of tests, when contemplating further research, and when contemplating recommendation for the application of tests in applied learning settings. In the following, I thus want to conclude by discussing these factors.

Desirability and Beneficial Long-Term Learning Effects of Tests

When focusing on the desirability of desirable difficulties, researchers at first typically contemplate potential beneficial effects of such difficult learning strategies for learners' later learning outcomes. This is necessary insofar as that only learning strategies that actually increase durable long-term learning outcomes can be recommended and should be applied. Learning strategies that are not beneficial in the long run—thus, those that are not desirable for durable learning outcomes—would not even merit further considerations. Hence, as a first step, it is valuable to investigate the long-term learning benefits of desirable difficulties. My conducted and here presented studies thereby mirrored the respective previous literature: For instance, the two papers investigating long-term effects of generation tasks indicated that these were either only beneficial for some learners (Reinhard et al., 2019; see Appendix B) or not beneficial at all (probably because our tasks were not difficult enough to require higher effort and to, in turn, elicit the beneficial cognitive processes; Wenzel & Reinhard, 2019, Study 1; see Appendix C). Although it was often shown that generation increases long-term

learning outcomes (for meta-analyses see: Bertsch et al., 2007; McCurdy et al., 2020), previous work also yielded that generation tasks did not continuously result in beneficial effects (see e.g., de Winstanley & Bjork, 2004; de Winstanley et al., 1996; Karpicke & Zaromb, 2010; Metcalfe & Kornell, 2007). Thus, it might be possible that generation tasks (which include a broad array of different activities and varying ways to stimulate self-generation) are not as robust or as easy to design and implement as assumed—especially compared to tests. Accordingly, some researchers previously argued that tests are more effective and more robust than generation tasks (e.g., Karpicke, 2017; Karpicke & Zaromb, 2010), some review articles explicitly recommended the application of tests compared to other learning strategies (e.g., Dunlosky et al., 2013; Pashler et al., 2007), and multiple meta-analyses highlight the strong, robust, and greatly generalizable benefits of tests (even compared to varying control conditions; e.g., Adesope et al., 2017; Agarwal et al., 2021; Pan & Rickard, 2018; Rowland, 2014; Schwieren et al., 2017; Sotola & Crede, 2021; Yang et al., 2021). These findings indicate that if lecturers—which often have only limited time and resources to implement further activities in their courses—had to choose among generations tasks, tests, other learning strategies, and further classroom activities, they should choose to use tests. In line with this consideration, my two conducted and here presented studies also showed that the application of short tests increased learners' learning outcomes compared to easier re-reading tasks—even for naturalistic, complex, and curricular learning materials that are actually applied in university courses (Wenzel & Reinhard, 2019, 2021b; see Appendix C and F). Because their application is also rather easy and not time-consuming (concerning both their preparation and their later implementation), tests can be strongly recommended for school or university settings: I would thus advise lecturers to use, for instance, the last 10 minutes of one or all of their course sessions to apply test questions concerning the contents of the respective lecture (ideally with feedback or corrections of errors; see e.g., Agarwal et al., 2021; S. Greving & Richter, 2018; Iwamoto et al., 2017; McDaniel et al., 2011; Pashler et

al., 2007; Yang et al., 2021). Positively, lecturers can thereby apply tests for a wide range of learning materials and contents, can implement varying test question formats, and can use one of multiple beneficial test application modes (see e.g., Adesope et al., 2017; Agarwal et al., 2021; Dunlosky et al., 2013; Pan & Rickard, 2018; Rowland, 2014; Yang et al., 2021).

Desirability and Intelligence as a Cognitive Prerequisite of Tests

Regarding potential cognitive prerequisites or boundary conditions for these desirable long-term learning effects of tests—which determine for whom tests should be (especially) applied and for whom not—recommendations are not as clear and cannot be given as confidently. In accordance, it was recently again stated that research investigating potential moderating effects of intelligence (and of cognitive abilities in general) are still scarce and that further work is still required (see e.g., Agarwal et al., 2021; Rummer, 2021). This request for further work is especially valuable given that the few existing studies resulted in varying findings (e.g., higher benefits of tests for more intelligent learners, at least concerning difficult items: Minear et al., 2018; higher benefits of tests for less intelligent learners: Brewer & Unsworth, 2012; no interaction between intelligence and the applied learning strategy: Karpicke et al., 2016; Robey, 2017). One of my here presented studies showed that learners' intelligence moderated the long-term learning effects of tests, insofar as that less intelligent learners did not benefit form tests, but averagely intelligent and especially higher intelligent learners did (Study 2, Wenzel & Reinhard, 2019; see Appendix C). Another one of my studies could not replicate these results and found no moderating effects of learners' intelligence (Wenzel & Reinhard, 2021b; see Appendix F). However, the first study can be considered more valid than this second study due to its higher sample size (N = 176 vs. N = 89) and higher power, so that my own findings would indicate a moderating effect of intelligence on the beneficial effects of tests. In contrast, Jonsson and colleagues (2021) later presented a study showing that cognitive abilities and intelligence did not act as boundary conditions for

the long-term learning effects of tests and that tests were beneficial for all learners independent of their intelligence. Notably, this would actually be the more desirable outcome, because then everyone would profit equally from tests, tests could be recommended without restrictions, and their benefits would be even more generalizable. However, as both evidence indicating that tests are beneficial for all learners independent of their intelligence as well as evidence indicating that tests are only beneficial for learners with average/higher intelligence exists, supporting one of these assumptions seems a bit premature. Additionally, validly interpreting and comparing these results is also hindered due to methodological differences between the existing work: Among others, the applied learning materials greatly differ concerning their level of difficulty and complexity (e.g., words/word pairs/associations: Brewer & Unsworth, 2012; Karpicke et al., 2016; Robey, 2017; Swahili and English/Swedish vocabulary: Minear et al., 2018; Jonsson et al., 2021; university textbook chapters: Wenzel & Reinhard, 2019, 2021b), concerning the measured intelligence estimate (e.g., speeded reasoning/processing speed: Karpicke et al., 2016; Wenzel & Reinhard, 2021b; fluid intelligence using matrices: Brewer & Unsworth, 2012; Minear et al., 2018; Robey, 2017; general cognitive ability/reasoning using verbal, numerical, and figural tasks: Wenzel & Reinhard, 2019; an accumulation of varying cognitive ability measures: Jonsson et al., 2021), and concerning the implemented delay (e.g., short delay/30-minutes: Karpicke et al., 2016; Robey, 2017; 1-day: Brewer & Unsworth, 2012; 2-days: Minear et al., 2018; 1-week: Jonsson et al., 2021; Wenzel & Reinhard, 2019, 2021b; 4-weeks: Jonsson et al., 2021). Thus, further work and replications are essential next steps for resolving these ambiguous findings. Only then can empirically well-grounded and valid recommendations for whom tests should be applied (or not) be given. Further studies could thereby focus more strongly on different theories, dimensions, or conceptualisations of intelligence and their potentially beneficial effects for the successful application of tests. Moreover, such future work should ideally be conducted in applied settings and should include complex and difficult learning materials as

well as longer test delays—such studies should be the most informative (see also: Agarwal et al., 2021; Rummer, 2021; Rummer & Schweppe, 2021). Until such further research exists, I argue that lecturers should still apply tests in their school or university courses. Even if at least average intelligence would be required to reap the benefits of tests, this should apply to most learners and the remaining learners would at least not suffer under the application of tests but would achieve similar learning outcomes as if they had used other learning strategies. Nonetheless, lecturers should keep in mind that some of their students might not benefit from the applied tests. Hence, if they would want to make sure that all their students would actually profit from tests, they could try to design the tests and the ensuing learning situations in such a way that learners with lower intelligence would also be able to reap their benefits. Previous literature suggested in this regard that, for instance, learners with lower abilities could benefit from having more time during the initial study opportunity (to be able to form better mental representations, to build an adequate knowledge base, and to ensure a higher understanding of the materials), from having more time to solve the difficult tasks (to increase the probability of being successful), or from working on desirable difficulties only later in the learning process (when they have already mastered some of the basic information and formed sufficient previous knowledge; see e.g., Kalyuga et al., 2003; Renkl et al., 2002; Rummer, 2021; Rummer & Schweppe, 2021; Schmidt-Weigand et al., 2009; Snow, 1989). Learners with lower intelligence could also profit from graded learning aids that support their learning process or from integrating worked examples and problem solving with fading and strategic prompts (see e.g., Hänze et al., 2010; Schmidt-Weigand et al., 2009). Lecturers could additionally use adaptive tests to ensure that the difficulty of the test is adequate for all learners independent of their cognitive abilities. That way, each learner would work on a test that is difficult enough to require increased effort but also not too difficult, thus ensuring that they can successfully work on the test without being cognitively overwhelmed (see also Endres & Renkl, 2015; Karpicke, 2017; Minear et al., 2018). Previous work accordingly

yielded that it is advantageous to use test questions with adaptive complexity and adaptive difficulty levels (see e.g., S. Greving et al., 2020; Heitmann et al., 2018). Adapting tests based on learners' cognitive load and cognitive demand ratings was also shown to increase the effectiveness of tests—even in a realistic university setting (Heitmann, Grund, et al., 2021; Heitmann, Obergassel, et al., 2021). Another similar approach would be to implement tests in an open-book instead of the until now described closed-book format: In contrast to closed-book tests where learners have to retrieval all answers to the test questions on their own, learners using open-book tests are allowed to consult the learning materials while answering the test questions. Both closed-book tests as well as open-book tests were continuously shown to lead to higher long-term learning outcomes compared to easier control tasks (e.g., Agarwal et al., 2008; Agarwal & Roediger, 2011; Arnold et al., 2021; Ebersbach, 2020; Rummer et al., 2019; Wenzel et al., 2021), but open-book tests might be especially beneficial for learners with lower intelligence. For instance, allowing learners to use the materials gives them the opportunity to build a more elaborate mental model, a more coherent mental representation, and a better situation model of the initial learning materials and generally increases their (initial) understanding of the to-be-learned information (see e.g., Roelle & Nückles, 2019; Rummer, 2021; Rummer & Schweppe, 2021; Waldeyer et al., 2020). This is generally beneficial but should be especially important and supporting for relatively less intelligent learners. Moreover, the application of open-book tests (including instructions to only use the materials when learners would not be able to retrieve the answer even after greatly increasing their effort and trying their best), could also lead to the needed balance between retrieval effort and retrieval success (see e.g., Roelle & Nückles, 2019; Rummer, 2021; Rummer & Schweppe, 2021; Waldeyer et al., 2020). Accordingly, switching flexibly between open-book and closed-book tests (see Waldeyer et al., 2020) might also be valuable to ensure that even learners with lower intelligence benefit from the application of tests.

Desirability and Negative Side-Effects of Tests

After thus contemplating the desirability of tests based on their long-term learning effects and on their generalizability for learners with varying levels of intelligence, researchers and lecturers should then focus on potentially unpleasant experiences, evaluations, or perceptions caused by tests. Thus, potential negative side-effects caused by tests are another dimension of desirability that should be considered. In line with previous work (e.g., Hinze & Rapp, 2014), my two studies included in this dissertation showed that low-stakes tests led to more negative evaluations of the situation and to higher stress and anxiety perceptions during learning (Wenzel & Reinhard, 2021a, 2021b; see Appendix D and F). Notably, these negative side-effects were not extremely strong but indicated that learners using tests had more undesirable experiences while learning compared to learners re-reading the same materials—even when the conducted tests was conducted as an extremely low-stakes and objectively stress-less situation. This is an important finding that should be clearly communicated, and that lecturers, learners, and researchers must try to deal with. However, because fitting research is still scarce (and needed; see e.g., Rummer, 2021), future research focusing thereon seems valuable. For instance, such work should ideally be conducted in actual school or university settings and should explicitly compare the negative side-effects caused by tests to potential negative side-effects caused by other learning activities or strategies other than the typically applied re-reading task. Because re-reading is as a very easy and fluent learning strategy that should be particularly low-stakes or stress-free, it might not be surprising that tests lead to comparably more negative side-effects. It has, however, until now not been investigated if tests would also result in more negative evaluations or higher stress and anxiety perceptions compared to other activities or strategies that learners are typically confronted with during their school or university classes (e.g., taking notes, participating in in-class-discussions, or giving presentations). Because this would influence

further considerations or potential recommendations, such further investigations are needed. Additionally, future work could also focus more closely on our previously presented assumption that the negative-side effects of tests might be less distinct for leaners with higher intelligence (Wenzel & Reinhard, 2021b; see Appendix F). Although we found no interaction between intelligence and the applied learning strategy (probably due to the small power), we could show that higher intelligence was generally linked to lower stress perceptions (see also: Goetz et al., 2007; LePine et al., 2004). Hence, it might be valuable to further investigate if higher intelligence could buffer the negative side-effects of tests (or not), because such more data could also shape considerations of these negative side-effects and the resulting recommendations. Until such further research exists, researchers and lecturers should assume that tests result generally and for alle learners in negative and undesirable side-effects and should thus try to reduce these. Therefore, lecturers could give more thorough descriptions and explanations of the beneficial effects of tests, which might decrease negative evaluations of this difficult learning strategy as well as learners' misconceptions that learning must feel easy to be effective or that effort is a sign of low ability (see e.g., Biwer et al., 2020; Bjork & Bjork, 2019; Rivers, 2021). Lecturers could also generally reward learners for increased effort or could try to reframe their appraisals or attributions of effort and failure (e.g., Abraham et al., 2019; Zepeda et al., 2020). However, introducing the benefits of desirable difficulties might be more challenging than one would suppose and only knowing about the effectiveness of tests might not be enough to ward learners form their negative side-effects or to increase their motivation (see e.g., Besken & Mulligan, 2014; Biwer et al., 2020; Bjork & Bjork, 2019; Rummer, 2021; Zepeda et al., 2020). Thus, lecturers and researchers should try to design tests (and the ensuing learning situations) in such ways that they are beneficial but do not lead to negative side-effects (or at least to fewer side-effects). For instance, lecturers could apply the afore-mentioned open-book tests instead of the traditionally applied closed-book test because learners were previously shown to prefer open-book to closed-book tests and reported lower

levels of anxiety after taking open-book instead of closed-book tests (Gharib et al., 2012). Additionally, previous work found that learners working with online tests performed at least equivalent in a later examination compared to learners working with paper-pencil tests but perceived online tests as less threatening (e.g., Cassady & Gridley, 2005; Dobson, 2008; see also: DeVaney, 2010; Szpunar et al., 2013; Yang et al., 2021). Given the recently risen importance of online courses and distance learning due to the increased amount of home-schooling during the COVID-19 pandemic, such considerations seem especially valuable and promising and should in any case be explored in more detail. Recent work from A. I. Wang and Tahir (2020) further implied that implementing tests with *Kahoot!* (a game-based learning platform mirroring an anonymous game show) might also be extremely desirable: More specifically, the authors described that tests applied with Kahoot! were beneficial for long-term learning outcomes but additionally increased learners' motivation, their enjoyment, their positive perceptions, and their satisfaction compared to paper-based tests and even reduced their anxiety and stress experiences (A. I. Wang & Tahir, 2020; see also: Iwamoto et al., 2017; Yabuno et al., 2019; for further potentially viable application modes see: Yang et al., 2021). A fully game-based test (embedded in a computer game treasure hunt) was also shown to lead to a better performance in a later examination as well as to lower anxiety experiences during the learning situation compared to a normal online test (Mavridis & Tsiatsos, 2017). Hence, using open-book tests, applying online tests, or implementing tests with varying creative or game-based application modes could be simultaneously desirable regarding learners' later leaning outcomes and regarding their perceptions and experiences while learning.

Desirability and Further Negative Consequences of Tests

Following these considerations on negative side-effects caused by tests, researchers and lecturers should then further contemplate later negative consequences caused indirectly

by tests via these side-effects. The in my dissertation presented findings can thereby be classified as especially undesirable, insofar as that they show that tests can trigger negative or stress-inducing experiences during learning that learners must suffer under and that are in turn even linked to increased (hypothetical) cheating (Wenzel & Reinhard, 2020; see Appendix E) and to a reduced effectiveness of the respective test (Wenzel & Reinhard, 2021b; see Appendix F). Given that cheating is an extremely unwanted behaviour that lecturers normally try to reduce or avoid and given that a less effective test would render the difficult and unpleasant learning situation less worthwhile and would run counter to the intention of using tests in the first place, such negative consequences should be considered when contemplating to use tests. However, because there is almost no other research simultaneously focusing on these important research issues, future work and replications are essential before valid and reliable implications and recommendations can be drawn from these results. This applies, for instance, to further research investigating if the here presented effects on academic cheating can even be replicated in realistic and applied settings (which include, among others, actual gains of cheating successfully but also actual costs of getting caught). Further work should also focus on ways to increase the by stress perceptions suppressed benefits of tests to realise their full potential (positively, our findings showed that tests were still more beneficial compared to re-reading even though their effectiveness was suppressed and though they were not as beneficial as they could have been—thereby again highlighting the strong and robust effects of tests). Thus, it seems to be important to communicate our findings concerning negative consequences indirectly caused by tests, so that lecturers and researchers are made aware of them and can thus try to cope with them or to directly prevent their occurrence. Fortunately, the previously described ways to reduce negative side-effects of tests should, in turn, also prevent (or at least reduce) the occurrence of negative consequences. Hence, applying tests that do not result in negative side-effects would not only optimize learners' experiences during learning but would thereby also avert these negative consequences.

The Desirability of Tests

Taking all these factors and the ensuing different dimensions of desirability of tests together—and given the current empirical data—I would advise lectures to nonetheless use tests in their respective courses. Firstly, tests are generally effective and can be classified as desirable concerning their later long-term learning benefits. Secondly, even though lecturers should keep in mind that average intelligence might act as a boundary condition for the successful application of tests and that in turn some of their students might not benefit from them, most learners should still be able to reap the benefits of tests. In addition, there seem to be multiple ways for lecturers to implement tests so that even learners with lower levels of intelligence can benefit from their application. Thirdly, even though tests can lead to acute negative side-effects during learning (like higher negative evaluations of the situation, more stress perceptions, and stronger anxiety experiences), these effects were fortunately not extremely strong and further work is still needed to ascertain if tests are even more unpleasant than typical classroom activities or learning strategies apart from re-reading. Positively, there appear to be varying ways to design tests to ensure that these lead to higher long-term learning outcomes without also resulting in negative side-effects. These procedures could be implemented and further investigated. Fourthly, even though tests can also indirectly lead to further negative consequences that can only be classified as extremely undesirable and that should be avoided, reducing the negative side-effects of tests by designing them adequately might also simultaneously prevent the occurrence of these negative consequences. Jointly, these considerations indicate that tests can still be classified as desirable and that lecturers can still be advised to apply them in their school or university courses. However, lecturers should thereby try to implement the ideas presented above on how to use and design tests so that all learners can reap their benefits equally without suffering under negative side-effects in the meantime and without also further triggering later occurring negative consequences.

Conclusion

All in all, this dissertation was conducted to take a more thorough look at the desirability of desirable difficulties—primarily concerning tests—and to focus on even more dimensions of desirability in addition to their effectiveness for later learning outcomes. I thereby ague that researchers and lecturers should not only consider beneficial long-term learning effects of tests when contemplating their application but should also consider cognitive prerequisites that might be required for learners to reap these benefits, negative side-effects caused by tests that arise immediately during learning, further negative consequences indirectly caused by tests, and finally also potential linkages among these factors. Thus, my dissertation was intended to increase the awareness of researchers and lecturers that it is valuable to simultaneously contemplate these factors, which might all determine the desirability of tests. I thereby also wanted to highlight my here presented research issues as well as the resulting findings and implications of my published papers as first steps for future research. Although my presented studies are not without limitations, although not all raised issues and questions could yet be tested conclusively, and although further replications are required for well-grounded recommendations and for supporting or refuting my described assumptions, it is nonetheless important to communicate the here presented issues. This way, learners, lectures, and researcher can gain a broader and more thorough understanding of tests (and desirable difficulties) and can hopefully receive helpful recommendation for their applications as well as stimulating ideas for the required further research.

References

Abel, R., & Hänze, M. (2019). Generating Causal Relations in Scientific Texts: The Long-Term Advantages of Successful Generation. *Frontiers in Psychology*, *10*, Article 199. https://doi.org/10.3389/fpsyg.2019.00199

Abín, A., Núñez, J. C., Rodríguez, C., Cueli, M., García, T., & Rosário, P. (2020). Predicting mathematics achievement in secondary education: The role of cognitive, motivational, and emotional variables. *Frontiers in Psychology*, *11,* Article 876. https://doi.org/10.3389/fpsyg.2020.00876

Abraham, D., McRae, K., & Mangels, J. A. (2019). "A" for Effort: Rewarding Effortful Retrieval Attempts Improves Learning From General Knowledge Errors in Women. *Frontiers in Psychology*, *10*, Article 1179. https://doi.org/10.3389/fpsyg.2019.01179

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659-701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K., D'Antonio, L., Roediger, H. L., III., McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition, 3*(3), 131–139. https://doi.org/10.1016/j.jarmac.2014.07.002

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*(7), 861–876. https://doi.org/10.1002/acp.1391

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval Practice Consistently Benefits Student Learning: A Systematic Review of Applied Research in Schools and

Classrooms. *Educational Psychology Review*, 1-45. https://doi.org/10.1007/s10648-021-09595-9

Agarwal, P. K., & Roediger, H. L., III. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory, 19*(8), 836–852. https://doi.org/10.1080/09658211.2011.613840

Agnew, R. (1992). Foundation for a general strain theory of crime and delinquency. *Criminology, 30*(1)*,* 47–88. https://doi.org/10.1111/j.1745-9125.1992.tb010 93.x

Alter, A. L., Oppenheimer, D. M., & Epley, N. (2013). Disfluency prompts analytic thinking—But not always greater accuracy: Response to. *Cognition*, *128*(2), 252-255. http://dx.doi.org/10.1016/j.cognition.2013.01.006

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, *136*(4), 569-576. https://doi.org/10.1037/0096-3445.136.4.569

Arnold, K. M., Eliseev, E. D., Stone, A. R., McDaniel, M. A., & Marsh, E. J. (2021). Two routes to the same place: learning from quick closed-book essays versus open-book essays. *Journal of Cognitive Psychology, 33*(1), 1-18. https://doi.org/10.1080/20445911.2021.1903011

Ashcraft, M. H., & Krause, J. A. (2007). Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review*, *14*(2), 243-248. https://doi.org/10.3758/BF03194059

Batsell, W. R., Jr., Perry, J. L., Hanley, E., & Hostetter, A. B. (2017). Ecological validity of the testing effect: The use of daily quizzes in introductory psychology. *Teaching of Psychology*, *44*(1), 18-23. https://doi.org/10.1177/0098628316677492

Baudson, T. G., & Preckel, F. (2015). mini-q: Intelligenzscreening in drei Minuten [mini-q: Intelligence screening in three minutes]. *Diagnostica*, *62*(3), 182-197. https://doi.org/10.1026/0012-1924/a000150.

Beilock, S. L. (2008). Math performance in stressful situations. *Current Directions in Psychological Science*, *17*(5), 339-343. https://doi.org/10.1111/j.14678721.2008.00602.x

Bertilsson, F., Stenlund, T., Wiklund-Hörnqvist, C., & Jonsson, B. (2021). Retrieval Practice: Beneficial for All Students or Moderated by Individual Differences?. *Psychology Learning & Teaching*, *20*(1), 21-39. https://doi.org/10.1177/1475725720973494

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201-210. https://doi.org/10.3758/BF03193441

Besken, M., & Mulligan, N. W. (2014). Perceptual fluency, auditory generation, and metamemory: Analyzing the perceptual fluency hypothesis in the auditory modality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(2), 429–440. https://doi.org/10.1037/a0034407

Biwer, F., de Bruin, A. B., Schreurs, S., & oude Egbrink, M. G. (2020). Future steps in teaching desirably difficult learning strategies: Reflections from the Study Smart program. *Journal of Applied Research in Memory and Cognition*, *9*(4), 439-446. https://doi.org/10.1016/j.jarmac.2020.07.006

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, J. R. Pomerantz (Eds.) & FABBS Foundation, *Psychology and the real*

world: Essays illustrating fundamental contributions to society (pp. 56–64). Worth Publishers.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). MIT Press.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). The MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes, Vol. 1. From learning theory to connectionist theory; Vol. 2. From learning processes to cognitive processes* (pp. 35–67). Lawrence Erlbaum Associates, Inc.

Bjork, R. A., & Bjork, E. L. (2019). Forgetting as the friend of learning: implications for teaching and self-regulated learning. *Advances in Physiology Education*, *43*(2), 164-167. https://doi.org/10.1152/advan.00001.2019

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, *9*(4), 475-479.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64,* 417-444. https://doi.org/10.1146/annurev-psych-113011-143823

Blake, K. R., & Gangestad, S. (2020). On attenuated interactions, measurement error, and statistical power: Guidelines for social and personality psychologists. *Personality and*

Social Psychology Bulletin, 46(12), 1702–1711. https://doi.org/10.1177/0146167220913363

Bornstein, M. H., Hahn, C. S., & Wolke, D. (2013). Systems and cascades in cognitive development and academic achievement. *Child Development*, *84*(1), 154-162. https://doi.org/10.1111/j.1467-8624.2012.01849.x

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language*, *66*(3), 407-415. https://doi.org/10.1016/j.jml.2011.12.009

Brimble, M., & Stevenson-Clarke, P. (2005). Perceptions of the prevalence and seriousness of academic dishonesty in Australian universities. *The Australian Educational Researcher, 32*(3)*,* 19–44. https://doi.org/10.1007/BF0321682

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin, 145*(11), 1029–1052. https://doi.org/10.1037/bul0000209

Bystritsky, A., & Kronemyer, D. (2014). Stress and anxiety: Counterpart elements of the stress/anxiety complex. *Psychiatric Clinics of North America, 37*(4), 489–518. https://doi.org/10.1016/j.psc.2014.08.002

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: support for the elaborative retrieval explanation of the testing effect. *Memory and Cognition, 34*(2), 268–276. https://doi.org/10.3758/BF03193405

Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review, 28*(2), 353–375. https://doi.org/10.1007/s10648-015-9311-9

Carrell, S. E., Malmstrom, F. V., & West, J. E. (2008). Peer effects in academic cheating. *Journal of Human Resources, 43*(1)*,* 173–207. https://doi.org/10.3368/jhr.43.1.173

Cassady, J. C. (2004a). The influence of cognitive test anxiety across the learning–testing cycle. *Learning and Instruction*, *14*(6), 569-592. https://doi.org/10.1016/j.learninstruc.2004.09.002

Cassady, J. C. (2004b). The impact of cognitive test anxiety on text comprehension and recall in the absence of external evaluative pressure. *Applied Cognitive Psychology*, *18*(3), 311-325. https://doi.org/10.1002/acp.968

Cassady, J. C., & Gridley, B. E. (2005). The effects of online formative and summative assessment on test anxiety and performance. *The Journal of Technology, Learning and Assessment*, *4*(1).

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Chen, I., & Chang, C. C. (2009). Cognitive load theory: An empirical study of anxiety and task performance in language learning. *Electronic Journal of Research in Educational Psychology, 7*(2), 729-746.

Chuderski, A. (2014). High intelligence prevents the negative impact of anxiety on working memory. *Cognition and Emotion*, *29*(7), 1197-1209. https://doi.org/10.1080/02699931.2014.969683

Clark, D. A., Crandall, J. R., & Robinson, D. H. (2018). Incentives and test anxiety may moderate the effect of retrieval on learning. *Learning and Individual Differences, 63,* 70–77. https://doi.org/10.1016/j.lindif.2018.03.001

Clark, D. A., & Svinicki, M. (2015). The effect of retrieval on post-task enjoyment of studying. *Educational Psychology Review, 27*(1), 51–67. https://doi.org/10.1007/s10648-014-9272-4

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Davis, S. F., Grover, C. A., Becker, A. H., & McGregor, L. N. (1992). Academic dishonesty: Prevalence, determinants, techniques, and punishments. *Teaching of Psychology, 19*(1)*,* 16–20. https://doi.org/10.1207/s1532 8023t op1901_3

DeVaney, T. A. (2010). Anxiety and attitude of graduate students in on-campus vs. online statistics courses. *Journal of Statistics Education*, *18*(1), 1-15. https://doi.org/10.1080/10691898.2010.11889472

de Winstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition, 32*(6), 945–955. https://doi.org/10.3758/BF03196872

de Winstanley, P. A., Bjork, E. L., & Bjork, R. A. (1996). Generation Effects and the Lack Thereof: The Role of Transfer-appropriate Processing. *Memory, 4*(1), 31–48. https://doi.org/10.1080/741940667

Dickhäuser, O., & Reinhard, M. A. (2006). Factors underlying expectancies of success and achievement: The influential roles of need for cognition and general or specific self-

concepts. *Journal of Personality and Social Psychology*, *90*(3), 490-500. https://doi.org/10.1037/0022-3514.90.3.490

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the bold (and the italicized): Effects of disfluency on educational outcomes. *Cognition*, *118*(1), 111-115. https://doi.org/10.1016/j.cognition.2010.09.012

Dobson, J. L. (2008). The use of formative online quizzes to enhance class preparation and scores on summative exams. *Advances in Physiology Education*, *32*(4), 297-302. https://doi.org/ 10.1152/advan.90162.2008.

Dobson, J. L., & Linderholm, T. (2015). The effect of selected "desirable difficulties" on the ability to recall anatomy information. *Anatomical Sciences Education*, *8*(5), 395-403. https://doi.org/10.1002/ase.1489

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58. https://doi.org/10.1177/1529100612453266

Ebersbach, M. (2020). Access to the learning material enhances learning by means of generating questions: Comparing open-and closed-book conditions. *Trends in Neuroscience and Education*, *19*, 100130. https://doi.org/10.1016/j.tine.2020.100130

Ebersbach, M., & Barzagar Nazari, K. (2020). Implementing distributed practice in statistics courses: Benefits for retention and transfer. *Journal of Applied Research in Memory and Cognition*, *9*(4), 532-541. https://doi.org/10.1016/j.jarmac.2020.08.014

Ebersbach, M., Feierabend, M., & Barzagar Nazari, K. (2020). Comparing the effects of generating questions, testing, and restudying on students' long-term recall in university

learning. *Applied Cognitive Psychology*, *34*(3), 724-736. https://doi.org/10.1002/acp.3639

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Edwards, J. A., & Templeton, A. (2005). The structure of perceived qualities of situations. *European Journal of Social Psychology*, *35*(6), 705-723. https://doi.org/10.1002/ejsp.271

Efklides, A., Papadaki, M., Papantoniou, G., & Kiosseoglou, G. (1997). Effects of cognitive ability and affect on school mathematics performance and feelings of difficulty. *The American Journal of Psychology*, *110*(2), 225-258. https://doi.org/10.2307/1423716

Eitel, A., Kühl, T., Scheiter, K., & Gerjets, P. (2014). Disfluency meets cognitive load in multimedia learning: Does harder-to-read mean better-to-understand? *Applied Cognitive Psychology, 28*(4), 488–501. https://doi.org/10.1002/acp.3004

Endler, N. S. (1997). Stress, Anxiety and coping: the multidimensional interaction model. *Canadian Psychology/Psychologie Canadienne*, *38*(3), 136-153. https://doi.org/10.1037/0708-5591.38.3.136

Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: An empirical investigation of retrieval practice in meaningful learning. *Frontiers in Psychology, 6,* Article 1054. https://doi.org/10.3389/fpsyg.2015.01054

Epel, E. S., Crosswell, A. D., Mayer, S. E., Prather, A. A., Slavich, G. M., Puterman, E., & Mendes, W. B. (2018). More than a feeling: A unified view of stress measurement for population science. *Frontiers in Neuroendocrinology*, *49*, 146-169. https://doi.org/10.1016/j.yfrne.2018.03.001

Eskenazi, M. A., & Nix, B. (2021). Individual differences in the desirable difficulty effect during lexical acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(1), 45–52. https://doi.org/10.1037/xlm0000809

Eysenck, M. W., & Calvo, M. G. (1992). Anxiety and performance: The processing efficiency theory. *Cognition & Emotion, 6*(6), 409-434. https://doi.org/10.1080/02699939208409696

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: attentional control theory. *Emotion*, *7*(2), 336-353. https://doi.org/10.1037/1528-3542.7.2.336

Fellman, D., Lincke, A., & Jonsson, B. (2020). Do Individual Differences in Cognition and Personality Predict Retrieval Practice Activities on MOOCs?. *Frontiers in Psychology*, *11*, Article 2076. https://doi.org/10.3389/fpsyg.2020.02076

Feraco, T., Casali, N., Tortora, C., Dal Bon, C., Accarrino, D., & Meneghetti, C. (2020). Using Mobile Devices in Teaching Large University Classes: How Does It Affect Exam Success?. *Frontiers in Psychology*, *11*, Article 1363. https://doi.org/10.3389/fpsyg.2020.01363

Fergusson, D. M., Horwood, L. J., & Ridder, E. M. (2005). Show me the child at seven II: Childhood intelligence and later outcomes in adolescence and young adulthood. *Journal of Child Psychology and Psychiatry*, *46*(8), 850-858. https://doi.org/10.1111/j.1469-7610.2005.01472.x

Fida, R., Tramontano, C., Paciello, M., Ghezzi, V., & Barbaranelli, C. (2018). Understanding the interplay among regulatory self-efficacy, moral disengagement, and academic cheating behaviour during vocational education: A three-wave study. *Journal of Business Ethics, 153*(3), 725–740. https://doi.org/10.1007/s1055 1-016-3373-6

Finn, K. V., & Frone, M. R. (2004). Academic performance and cheating: Moderating role of school identification and self-efficacy. *The Journal of Educational Research, 97*(3)*,* 115–121. https://doi.org/10.3200/JOER.97.3.115-121

Freiburger, T. L., Romain, D. M., Randol, B. M., & Marcum, C. D. (2017). Cheating behaviors among undergraduate college students: Results from a factorial survey. *Journal of Criminal Justice Education, 28*(2)*,* 222–247. https://doi.org/10.1080/10511253.2016.12030 10

García, T., Rodríguez, C., Betts, L., Areces, D., & González-Castro, P. (2016). How affective-motivational variables and approaches to learning relate to mathematics achievement in upper elementary levels. *Learning and Individual Differences, 49,* 25-31. https://doi.org/10.1016/j.lindif.2016.05.021

Gardiner, J. M., & Hampton, J. A. (1985). Semantic memory and the generation effect: Some tests of the lexical activation hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 732–741. https://doi.org/10.1037/0278-7393.11.1-4.732

Gharib, A., Phillips, W., & Mathew, N. (2012). Cheat Sheet or Open-Book? A Comparison of the Effects of Exam Types on Performance, Retention, and Anxiety. *Psychology Research*, *2*(8), 469-478. https://doi.org/10.17265/2159-5542/2012.08.044

Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological Science, 20*(3)*,* 393–398. https://doi.org/10.1111/j.1467-9280.2009.02306.x

Goetz, T., Preckel, F., Pekrun, R., & Hall, N. C. (2007). Emotional experiences during test taking: Does cognitive ability make a difference? *Learning and Individual Differences*, *17*(1), 3-16. https://doi.org/10.1016/j.lindif.2006.12.002

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24*(1), 79–132. https://doi.org/10.1016/S0160-2896(97)90014-3

Greving, C. E., & Richter, T. (2021). Beyond the Distributed Practice Effect: Is Distributed Learning Also Effective for Learning With Non-repeated Text Materials?. *Frontiers in Psychology*, *12,* Article 685245. https://doi.org/10.3389/fpsyg.2021.685245

Greving, S., Lenhard, W., & Richter, T. (2020). Adaptive retrieval practice with multiple-choice questions in the university classroom. *Journal of Computer Assisted Learning*, *36*(6), 799-809. https://doi.org/10.1111/jcal.12445

Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: Retrievability and question format matter. *Frontiers in Psychology, 9,* Article 2412. https://doi.org/10.3389/fpsyg.2018.02412

Haines, V. J., Diekhoff, G. M., LaBeff, E. E., & Clark, R. E. (1986). College cheating: Immaturity, lack of commitment, and the neutralizing attitude. *Research in Higher Education, 25*(4)*,* 342–354. https://doi.org/10.1007/BF00992130

Hänze, M., Schmidt-Weigand, F., & Stäudel, L. (2010). Gestufte Lernhilfen [Graded learning aids]. *Individuelle Förderung durch Innere Differenzierung. Ein Praxishandbuch für Lehrerinnen und Lehrer der Sekundarstufe II*, 63-73.

Heitmann, S., Grund, A., Berthold, K., Fries, S., & Roelle, J. (2018). Testing is more desirable when it is adaptive and still desirable when compared to note-taking. *Frontiers in Psychology, 9,* Article 2596. https://doi.org/10.3389/fpsyg.2018.02596

Heitmann, S., Grund, A., Fries, S., Berthold, K., & Roelle, J. (2021). The quizzing effect depends on hope of success and can be optimized by cognitive load-based

adaptation. *Learning and Instruction*, *77*, 101526. https://doi.org/10.1016/j.learninstruc.2021.101526

Heitmann, S., Obergassel, N., Fries, S., Grund, A., Berthold, K., & Roelle, J. (2021). Adaptive Practice Quizzing in a University Lecture: A Pre-Registered Field Experiment. *Journal of Applied Research in Memory and Cognition*. https://doi.org/10.1016/j.jarmac.2021.07.008

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, *58*(1), 47–77. https://doi.org/10.3102/00346543058001047

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology*, *28*(4), 597-606. https://doi.org/10.1002/acp.3032

Hirshman, E., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 484–494. https://doi.org/10.1037/0278-7393.14.3.484

Hobfoll, S. E. (1989). Conservation of resources: A new attempt at conceptualizing stress. *American Psychologist*, *44*(3), 513-524. https://doi.org/10.1037/0003-066X.44.3.513

Hoffman, E., & Spitzer, M. L. (1985). Entitlements, rights, and fairness: An experimental examination of subjects' concepts of distributive justice. *The Journal of Legal Studies, 14*(2), 259–297. https://doi.org/10.1086/467773

Houser, D., Vetter, S., & Winter, J. (2012). Fairness and cheating. *European Economic Review, 56*(8)*,* 1645–1655. https://doi.org/10.1016/j.euroe corev.2012.08.001

Iwamoto, D. H., Hargis, J., Taitano, E. J., & Vuong, K. (2017). Analyzing the efficacy of the testing effect using KahootTM on student performance. *Turkish Online Journal of Distance Education*, *18*(2), 80-93. https://doi.org/10.17718/tojde.306561

Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied, 22*(3), 305–318. https://doi.org/10.1037/xap0000087

Johnson, D. R., & Gronlund, S. D. (2009). Individuals lower in working memory capacity are particularly vulnerable to anxiety's disruptive effect in performance. *Anxiety, Stress, & Coping, 22*(2)*,* 201-213. https://doi.org/10.1080/10615800802291277

Jonsson, B., Wiklund-Hörnqvist, C., Stenlund, T., Andersson, M., & Nyberg, L. (2021). A learning method for all: The testing effect is independent of cognitive ability. *Journal of Educational Psychology, 113*(5), 972–985. https://doi.org/10.1037/edu0000627

Kaiser, I., Mayer, J., & Malai, D. (2018). Self-Generation in the Context of Inquiry-Based Learning. *Frontiers in Psychology*, *9*, Article 2440. https://doi.org/10.3389/fpsyg.2018.02440

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*(1), 23-31. https://doi.org/10.1207/S15326985EP3801_4

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology*, *93*(3), 579-588. https://doi.org/10.1037/0022-0663.93.3.579

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*(4-5), 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D. (2017). Retrieval-Based Learning: A Decade of Progress. In J. T. WixtedJ. H. Byrne (Ed.), *Cognitive Psychology of Memory, Vol. 2 of Learning and Memory: A Comprehensive Reference* (J. H. Byrne, Series Ed.) (pp. 487-514). Academic Press. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*(2), 317-326. https://doi.org/10.1007/s10648-015-9309-3

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborate studying with concept mapping. *Science, 331*(6018), 772–775. https://doi.org/10.1126/science.1199327

Karpicke, J. D., Blunt, J. R., & Smith, M. A. (2016). Retrieval-based learning: Positive effects of retrieval practice in elementary school children. *Frontiers in Psychology, 7,* Article 350. https://doi.org/10.3389/fpsyg.2016.00350

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own?. *Memory*, *17*(4), 471-479. https://doi.org/10.1080/09658210802647009

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 61, pp. 237–284). Elsevier Academic Press

Karpicke, J. D., & Roediger, H. L., III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, 33(4), 704–719. https://doi.org/10.1037/0278-7393.33.4.704

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62*(3), 227–239. https://doi.org/10.1016/j.jml.2009.11.010

Kausar, R. (2010). Perceived Stress, Academic Workloads and Use of Coping Strategies by University Students. *Journal of Behavioural Sciences*, *20*(1), 31-45.

Keeley, J., Zayac, R., & Correia, C. (2008). Curvilinear relationships between statistics anxiety and performance among undergraduate students: Evidence for optimal anxiety. *Statistics Education Research Journal, 7*(1), 4–15.

Khan, M. J., Altaf, S., & Kausar, H. (2013). Effect of Perceived Academic Stress on Students' Performance. *FWU Journal of Social Sciences*, *7*(2), 146-151.

Khanna, M. M. (2015). Ungraded pop quizzes: test-enhanced learning without all the anxiety. *Teaching of Psychology*, *42*(2), 174-178. https://doi.org/ 10.1177/0098628315573144

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology, 115,* Article 101237. https://doi.org/10.1016/j.cogpsych.2019.101237

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory and Cognition, 34*(5), 959-72. https://doi.org/10.3758/BF03193244

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 989–998. https://doi.org/10.1037/a0015729

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science*, *22*(6), 787-794. https://doi.org/10.1177/0956797611407929

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 183–215). Elsevier Academic Press.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology*, *86*(1), 148-161. https://doi.org/10.1037/0022-3514.86.1.148

Kurebayashi, L. F. S., Do Prado, J. M., & Da Silva, M. J. P. (2012). Correlations between stress and anxiety levels in nursing students. *Journal of Nursing Education and Practice*, *2*(3), 128-124. https://doi.org/10.5430/jnep.v2n3p128

Lazarus, R. S. (1990). Theory-based stress measurement. *Psychological Inquiry*, *1*(1), 3-13. https://doi.org/10.1207/s15327965pli0101_1

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. Springer Publishing Company.

Lazarus, R. S., & Folkman, S. (1987). Transactional theory and research on emotions and coping. *European Journal of Personality*, *1*(3), 141-169. https://doi.org/10.1002/per.2410010304

Lechuga, M. T., Ortega-Tudela, J. M., & Gómez-Ariza, C. J. (2015). Further evidence that concept mapping is not better than repeated retrieval as a tool for learning from texts. *Learning and Instruction, 40*, 61–68. https://doi.org/10.1016/j.learninstruc.2015.08.002

Lehmann, J., Goussios, C., & Seufert, T. (2016). Working memory capacity and disfluency effect: An aptitude-treatment-interaction study. *Metacognition and Learning*, *11*(1), 89-105. https://doi.org/10.1007/s11409-015-9149-z

Leiner, J. E. M., Scherndl, T., & Ortner, T. M. (2018). How Do Men and Women Perceive a High-Stakes Test Situation?. *Frontiers in Psychology*, *9*, Article 2216. https://doi.org/10.3389/fpsyg.2018.02216

LePine, J. A., LePine, M. A., & Jackson, C. L. (2004). Challenge and hindrance stress: relationships with exhaustion, motivation to learn, and learning performance. *Journal of Applied Psychology*, *89*(5), 883-891. https://doi.org/10.1037/0021-9010.89.5.883

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R (IST 2000 R)* [Intelligence-Structure-Test 2000 R (IST 2000 R)]. *Manual (2. erweiterte und überarbeitete Auflage.).* Hogrefe.

Lipowsky, F., Richter, T., Borromeo-Ferri, R., Ebersbach, M., & Hänze, M. (2015). Wünschenswerte Erschwernisse beim Lernen [Desirable difficulties during learning]. *Schulpädagogik heute*, *6*(11), 1-10. https://doi.org/10.17170/kobra-202102183286

Maass, J. K., & Pavlik, P. I., Jr. (2016). Modeling the Influence of Format and Depth during Effortful Retrieval Practice. *International Educational Data Mining Society*.

Mariss, A.*, Wenzel, K.*, Grünberg, C., & Reinhard, M.-A. (2022). Who wants to learn harder? The relationship between conservatism and liberalism, desirable difficulties, and academic learning. *Social Psychology of Education, 25,* 209-248. https://doi.org/10.1007/s11218-021-09681-4, * shared first-authorship.

Marshall, M. A., & Brown, J. D. (2004). Expectations and realizations: The role of expectancies in achievement settings. *Motivation and Emotion*, *28*(4), 347-361. https://doi.org/10.1007/s11031-004-2388-y

Mavridis, A., & Tsiatsos, T. (2017). Game-based assessment: Investigating the impact on test anxiety and exam performance. *Journal of Computer Assisted Learning*, *33*(2), 137-150. https://doi.org/10.1111/jcal.12170

McCabe, D. L. (1992). The influence of situational ethics on cheating among college students. *Sociological Inquiry, 62*(3)*,* 365–374. https://doi.org/10.1111/j.1475-682X.1992.tb00287.x

McCabe, D. L., Treviño, L. K., & Butterfield, K. D. (2001). Cheating in academic institutions: A decade of research. *Ethics and Behavior, 1*(3)1*,* 219–232. https://doi.org/10.1207/S15327019E B11032

McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. (2020). Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review, 27*(6), 1139–1165. https://doi.org/10.3758/s13423-020-01762-3

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. Journal of Educational Psychology, *103*(2), 399–414. https://doi.org/10.1037/a0021782

McDaniel, M. A., & Butler, A. C. (2011). A contextual framework for understanding when difficulties are desirable. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 175–198). Psychology Press.

McDaniel, M. A., Hines, R. J., & Guynn, M. J. (2002). When text difficulty benefits less-skilled readers. *Journal of Memory and Language*, *46*(3), 544-561. https://doi.org/10.1006/jmla.2001.2819

McDaniel, M. A., Roediger, H. L., III., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*(2), 200-206. https://doi.org/10.3758/BF03194052

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L., III. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, *27*(3), 360-372. https://doi.org/10.1002/acp.2914

McGrath, J. E. (1970). Social and Psychological Factors in Stress. Holt, Rinehart, and Winston.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*(1), 1-43. https://doi.org/10.1207/s1532690xci1401_1

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review, 14*(2), 225–229. https://doi.org/10.3758/BF03194056

Miele, D. B., Finn, B., & Molden, D. C. (2011). Does easily learned mean easily remembered? It depends on your beliefs about intelligence. *Psychological Science, 22*(3), 320–324. https://doi.org/10.1177/0956797610397954

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(9), 1474-1486. https://doi.org/10.1037/xlm0000486

Mok, W. S. Y., & Chan, W. W. L. (2016). How do tests and summary writing tasks enhance long-term retention of students with different levels of test anxiety? *Instructional Science*, *44*(6), 567-581. https://doi.org/10.1007/s11251-016-9393-x

Moreno, R., Reisslein, M., & Ozogul, G. (2009). Optimizing worked-example instruction in electrical engineering: The role of fading and feedback during problem-solving practice. *Journal of Engineering Education*, *98*(1), 83-92. https://doi.org/10.1002/j.2168-9830.2009.tb01007.x

Muenks, K., Miele, D. B., & Wigfield, A. (2016). How students' perceptions of the source of effort influence their ability evaluations of other students. *Journal of Educational Psychology, 108*(3), 438–454. https://doi.org/10.1037/edu0000068

Naveh-Benjamin, M. (1991). A comparison of training programs intended for different types of test-anxious students: Further support for an information-processing model. *Journal of Educational Psychology, 83(1)*, 134–139. https://doi.org/10.1037/0022-0663.83.1.134

Nemeth, L., Werker, K., Arend, J., & Lipowsky, F. (2021). Fostering the acquisition of subtraction strategies with interleaved practice: An intervention study with German

third graders. *Learning and Instruction*, *71*, 101354. https://doi.org/10.1016/j.learninstruc.2020.101354

Nyroos, M., Schéle, I., & Wiklund-Hörnqvist, C. (2016). Implementing Test Enhanced Learning: Swedish Teacher Students' Perception of Quizzing. *International Journal of Higher Education*, *5*(4), 1-12. https://doi.org/10.5430/ijhe.v5n4p1

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H. M. (2005). Working Memory and Intelligence—Their Correlation and Their Relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 61-65. https://doi.org/10.1037/0033-2909.131.1.61

Olafson, L., Schraw, G., Nadelson, L., Nadelson, S., & Kehrwald, N. (2013). Exploring the judgment–action gap: College students and academic dishonesty. *Ethics and Behavior, 23*(2)*,* 148–162. https://doi.org/10.1080/10508 422.2012.714247

O'Neil, J. H., Spielberger, C. D., & Hansen, D. N. (1969). Effects of state anxiety and task difficulty on computer-assisted learning. *Journal of Educational Psychology*, *60*(5), 343-350. https://doi.org/10.1037/h0028323

Oppenheimer, D. M., & Alter, A. L. (2014). The search for moderators in disfluency research. *Applied Cognitive Psychology, 28*(4), 502–504. https://doi.org/10.1002/acp.3023

Ott, M. K. (2017). *Testangst und deren Zusammenhang mit Testleistung: Effekte von Messzeitpunkt, Instruktion und funktionaler Bewertung der Angst [Test anxiety and its relationship to test performance: effects of time of measurement, instruction and functional assessment of anxiety]* (Doctoral dissertation, Justus-Liebig-Universität Giessen).

Owens, M., Stevenson, J., Hadwin, J. A., & Norgate, R. (2014). When does anxiety help or hinder cognitive test performance? The role of working memory capacity. *British Journal of Psychology*, *105*(1), 92-101. https://doi.org/10.1111/bjop.12009

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*(7), 710–756. https://doi.org/10.1037/bul0000151

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). Organizing Instruction and Study to Improve Student Learning. IES Practice Guide. NCER 2007-2004. *National Center for Education Research*.

Paternoster, R., McGloin, J. M., Nguyen, H., & Thomas, K. J. (2013). The causal impact of exposure to deviant peers: An experimental investigation. *Journal of Research in Crime and Delinquency, 50*(4)*,* 476–503. https://doi.org/10.1177/0022427812444274

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General, 143*(2), 644–667. https://doi.org/10.1037/a0033194

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437-447. https://doi.org/10.1016/j.jml.2009.01.004

Rauthmann, J. F. (2012). You say the party is dull, I say it is lively: A componential approach to how situations are perceived to disentangle perceiver, situation, and perceiver× situation variance. *Social Psychological and Personality Science*, *3*(5), 519-528. https://doi.org/10.1177/1948550611427609

Rawson, K. A. (2015). The status of the testing effect for complex materials: Still a winner. *Educational Psychology Review, 27*(2), 327–331. https://doi.org/10.1007/s10648-015-9308-4

Reeve, C. L., Bonaccio, S., & Winford, E. C. (2014). Cognitive ability, exam-related emotions and exam performance: A field study in a college setting. *Contemporary Educational Psychology, 39*(2), 124–133. https://doi.org/10.1016/j.cedpsych.2014.03.001

Reinhard, M.-A., Weissgerber, S. C., & Wenzel, K.* (2019). Performance expectancies moderate the effectiveness of worked-examples and problem-solving over time. *Frontiers in Psychology, 10*, Article 1623. https://doi.org/10.3389/fpsyg.2019.01623 *shared first-authorship.

Reinhardt, N., Wenzel, K., & Reinhard, M.-A. (2019). Am I responsible for my learning success? A study about the correlation between locus of control and attitudes towards and self-reported use of desirable difficulties. *Journal of Psychological and Educational Research, 27*(1), 7–24.

Renkl, A., Atkinson, R. K., Maier, U. H., & Staley, R. (2002). From example study to problem solving: Smooth transitions help learning. *Journal of Experimental Education, 70*(4), 293–315. https://doi.org/10.1080/00220970209599510

Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). Linking cognitive science to education: Generation and interleaving effects. In *Proceedings of the twenty-seventh annual conference of the Cognitive Science Society* (pp. 1850-1855). Erlbaum.

Rivers, M. L. (2021). Metacognition about practice testing: A review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educational Psychology Review, 33*(3), 823–862. https://doi.org/10.1007/s10648-020-09578-2

Robey, A. M. (2017). *The benefits of testing: Individual differences based on student factors* (Doctoral dissertation, University of Maryland).

Roediger, H. L., III., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382-395. https://doi.org/10.1037/a0026252

Roediger, H. L., III., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., III., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3)*,* 249-255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roelle, J., & Nückles, M. (2019). Generative learning versus retrieval practice in learning from text: The cohesion and elaboration of the text matters. *Journal of Educational Psychology, 111*(8), 1341–1361. https://doi.org/10.1037/edu0000345

Rost, D. H., & Wild, K. P. (1994). Cheating and achievement-avoidance at school: Components and assessment. *British Journal of Educational Psychology, 64*(1)*,* 119–132. https://doi.org/10.1111/j.2044-8279.1994.tb010 89.x

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118-137. https://doi.org/10.1016/j.intell.2015.09.002

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Rummer, R. (2021). Der Testungseffekt beim Lernen mit Texten [The testing effect when learning with texts]. *Psychologische Rundschau*. *72*(4), 259–272. https://doi.org/10.1026/0033-3042/a000518

Rummer, R., & Schweppe, J. (2021). Komplexität und der Testungseffekt: Die mögliche Bedeutung der Verständnissicherung für den Nutzen von Abrufübung bei komplexem Lernmaterial [Complexity and the testing effect: The possible importance of securing comprehension for the benefits of retreival practice for complex learning materials]. *Unterrichtswissenschaft*, 1-16. https://doi.org/10.1007/s42010-021-00137-4

Rummer, R., Schweppe, J., Gerst, K., & Wagner, S. (2017). Is testing a more effective learning strategy than note-taking? *Journal of Experimental Psychology: Applied, 23*(3), 293–300. https://doi.org/10.1037/xap0000134

Rummer, R., Schweppe, J., & Schwede, A. (2019). Open-book versus closed-book tests in university classes: A field experiment. *Frontiers in Psychology*, *10*, Article 463. https://doi.org/10.3389/fpsyg.2019.00463

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology*, *46*(4), 929-938. https://doi.org/10.1037/0022-3514.46.4.929

Sarason, I. G., & Sarason, B. R. (1990). Test anxiety. In *Handbook of Social and Evaluation Anxiety* (pp. 475-495). Springer. https://doi.org/10.1007/978-1-4899-2504-6_16

Schab, F. (1991). Schooling without learning: Thirty years of cheating in high school. *Adolescence, 26*(104), 839–847.

Schindler, J., Schindler, S., & Reinhard, M.-A. (2019). Effectiveness of self-generation during learning is dependent on individual differences in need for cognition. *Frontline Learning Research, 7*(2), 23-39. https://doi.org/10.14786/flr.v7i2.407

Schmidt-Weigand, F., Hänze, M., & Wodzinski, R. (2009). Complex problem solving and worked examples: The role of prompting strategic behavior and fading-in solution steps. *Zeitschrift für Pädagogische Psychologie / German Journal of Educational Psychology, 23*(2), 129–138. https://doi.org/10.1024/1010-0652.23.2.129

Schunk, H. D., & Gaa, J. P. (1981). Goal-setting influence on learning and self-evaluation. *The Journal of Classroom Interaction*, *16*(2), 38-44.

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: A meta-analytic perspective. *Psychology Learning & Teaching, 16*(2), 179–196. https://doi.org/10.1177/1475725717695149

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research*, *4*(1), 27-41. https://doi.org/10.1080/08917779108248762

Snow, R. E. (1989). Aptitude-treatment interaction as a framework for research on individual differences in learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences: Advances in theory and research* (pp. 13–59). W H Freeman/Times Books/ Henry Holt & Co.

Sotardi, V. A., Bosch, J., & Brogt, E. (2020). Multidimensional influences of anxiety and assessment type on task performance. *Social Psychology of Education*, *23*(2), 499-522. https://doi.org/10.1007/s11218-019-09508-3

Sotola, L. K., & Crede, M. (2021). Regarding Class Quizzes: A Meta-Analytic Synthesis of Studies on the Relationship between Frequent Low-Stakes Testing and Class Performance. *Educational Psychology Review*, *33*(2), 407-426. https://doi.org/10.1007/s10648-020-09563-9.

Stadler, M., Becker, N., Gödker, M., Leutner, D., & Greiff, S. (2015). Complex problem solving and intelligence: A meta-analysis. *Intelligence, 53,* 92–101. https://doi.org/10.1016/j.intell.2015.09.005

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*(4), 360–407. https://doi.org/10.1598/RRQ.21.4.1

Steininger, M., Johnson, R. E., & Kirts, D. K. (1964). Cheating on college examinations as a function of situationally aroused anxiety and hostility. *Journal of Educational Psychology, 55*(6)*,* 317–324. https://doi.org/10.1037/h0042396

Stern, E. (2015). Intelligence, prior knowledge, and learning. *International Encyclopedia of the Social and Behavioral Sciences (2nd ed, Vol. 12, pp. 323–328). Elsevier*. https://doi.org/10.1016/B978-0-08-097086-8.92017-8

Stern, E. (2017). Individual differences in the learning potential of human beings. *npj Science of Learning*, *2*(1), 2-7. https://doi.org/10.1038/s41539-016-0003-0

Sternberg, R. J. (1997). The concept of intelligence and its role in lifelong learning and success. *American Psychologist, 52*(10), 1030–1037. https://doi.org/10.1037/0003-066X.52.10.1030

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, *35*(5), 401-426. https://doi.org/10.1016/j.intell.2006.09.004

Struthers, C. W., Perry, R. P., & Menec, V. H. (2000). An examination of the relationship among academic stress, coping, motivation, and performance in college. *Research in Higher Education*, *41*(5), 581-592. https://doi.org/10.1023/A:1007094931292

Sung, Y. T., Chao, T. Y., & Tseng, F. L. (2016). Reexamining the relationship between test anxiety and learning achievement: An individual-differences perspective. *Contemporary Educational Psychology*, *46*, 241-252. https://doi.org/10.1016/j.cedpsych.2016.07.001

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, *110*(16), 6313-6317. https://doi.org/10.1073/pnas.1221764110

Tobias, S. (1984). Test Anxiety: Cognitive Interference or Inadequate Preparation? *in Annual Meeting of the American Educational Research Association*. April 23–27, 1984; New Orleans, LA

Tse, C. S., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *Journal of Experimental Psychology: Applied*, *18*(3), 253-264. https://doi.org/10.1037/a0029190

Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(6), 607-617. https://doi.org/10.1037/0278-7393.5.6.607

Unsworth, N. (2010). On the division of working memory and long-term memory and their relation to intelligence: A latent variable approach. *Acta Psychologica*, *134*(1), 16-28. https://doi.org/10.1016/j.actpsy.2009.11.010

Unsworth, N. (2019). Individual differences in long-term memory. *Psychological Bulletin, 145*(1), 79–139. https://doi.org/10.1037/bul0000176

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology, 71,* 1–26. https://doi.org/10.1016/j.cogpsych.2014.01.003

van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: the testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*(2), 247-264. https://doi.org/10.1007/s10648-015-9310-x

Waldeyer, J., Heitmann, S., Moning, J., & Roelle, J. (2020). Can generative learning tasks be optimized by incorporation of retrieval practice? *Journal of Applied Research in Memory and Cognition, 9*(3), 355-369. https://doi.org/10.1016/j.jarmac.2020.05.001

Wang, A. I., & Tahir, R. (2020). The effect of using Kahoot! for learning–A literature review. *Computers & Education*, *149*, 103818. https://doi.org/10.1016/j.compedu.2020.103818

Wang, T., Ren, X., & Schweizer, K. (2017). Learning and retrieval processes predict fluid intelligence over and above working memory. *Intelligence*, *61*, 29-36. https://doi.org/10.1016/j.intell.2016.12.005

Weissgerber, S. C., & Reinhard, M.-A. (2017). Is disfluency desirable for learning? *Learning and Instruction*, *49*, 199-217. https://doi.org/10.1016/j.learninstruc.2017.02.004

Weissgerber, S. C., & Reinhard, M.-A. (2018). Pilot Study on the Relationship of Test Anxiety to Utilizing Self-Testing in Self-Regulated Learning. *International Journal of Psychological Studies, 10*(4)*, 95-109.* https://doi.org/*10.5539/ijps.v10n4p95*

Weissgerber, S. C., Reinhard, M.-A., & Schindler, S. (2016). Study harder? The relationship of achievement goals to attitudes and self-reported use of desirable difficulties in self-regulated learning. *Journal of Psychological and Educational Research*, *24*(1), 42–60.

Weissgerber, S. C., Reinhard, M.-A., & Schindler, S. (2018). Learning the hard way: Need for cognition influences attitudes towards and self-reported use of desirable learning difficulties. *Educational Psychology: An International Journal of Experimental Educational Psychology*, *38*(2), 176–202. doi:10.1080/01443410.2017.1387644

Wenzel, K., & Reinhard, M.-A. (2019). Relatively unintelligent individuals do not benefit from intentionally hindered learning: The role of desirable difficulties. *Intelligence, 77,* 101405. https://doi.org/10.1016/j.intell.2019.101405

Wenzel, K., & Reinhard, M.-A. (2020). Tests and academic cheating: do learning tasks influence cheating by way of negative evaluations?. *Social Psychology of Education, 23*(3), 721–753. https://doi.org/10.1007/s11218-020-09556-0

Wenzel, K., & Reinhard, M.-A. (2021a). Does the end justify the means? Learning tests lead to more negative evaluations and to more stress experiences. *Learning and Motivation, 73*, Article 101706. https://doi.org/10.1016/j.lmot.2020.101706

Wenzel, K., & Reinhard, M.-A. (2021b). Learning with a double-edged sword? Beneficial and detrimental effects of learning tests—taking a first look at linkages among tests, later learning outcomes, stress perceptions, and intelligence. *Frontiers in Psychology*, *12*, Article 693585. https://doi.org/10.3389/fpsyg.2021.693585

Wenzel, K., Schweppe, J., & Rummer, R. (2022*).* Are open-book tests still as effective as closed-book tests even after a delay of two weeks? *Applied Cognitive Psychology*, *36*(3), 699-707. https://doi.org/10.1002/acp.3943

Whitley, B. E. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education, 39*(3)*,* 235–274. https://doi.org/10.1023/A:1018724900565

Wine, J. (1971). Test anxiety and direction of attention. *Psychological Bulletin*, *76*(2), 92–104. https://doi.org/10.1037/10037-000

Wowra, S. A. (2007). Moral identities, social anxiety, and academic dishonesty among American college students. *Ethics and Behavior, 17*(3)*,* 303–321. https://doi.org/10.1080/10508420701519312

Yabuno, K., Luong, E., & Shaffer, J. F. (2019). Comparison of Traditional and Gamified Student Response Systems in an Undergraduate Human Anatomy Course. *HAPS Educator*, *23*(1), 29-36. https://doi.org/10.21692/haps.2019.001

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin, 147*(4), 399–435. https://doi.org/10.1037/bul0000309

Yang, C., Sun, B., Potts, R., Yu, R., Luo, L., & Shanks, D. R. (2020). Do working memory capacity and test anxiety modulate the beneficial effects of testing on new learning? *Journal of Experimental Psychology: Applied, 26*(4), 724–738. https://doi.org/10.1037/xap0000278

Zepeda, C. D., Martin, R. S., & Butler, A. C. (2020). Motivational strategies to engage learners in desirable difficulties. *Journal of Applied Research in Memory and Cognition*, *9*(4), 468-474. https://doi.org/10.1016/j.jarmac.2020.08.007

Ziegler, E., & Stern, E. (2014). Delayed benefits of learning elementary algebraic transformations through contrasted comparisons. *Learning and Instruction*, *33,* 131-146. https://doi.org/10.1016/j.learninstruc.2014.04.006

STAMENT OF ORIGINALITY

Universität Kassel, Fachbereich Humanwissenschaften

Erklärung zur kumulativen Dissertation im Promotionsfach Psychologie

Erklärung über den Eigenanteil an den publizierten Artikeln innerhalb der Dissertationsschrift und eidesstattliche Versicherung.

Allgemeine Angaben

Name: Wenzel, Kristin

Institut: Institut für Psychologie, Universität Kassel

Thema der Dissertation: „*Taking a closer look at the desirability of desirable difficulties—additionally focusing on required prerequisites, negative side-effects, and negative consequences*“

Erklärung gemäß § 8 der Allgemeinen Bestimmungen für Promotionen der Universität Kassel vom 14.07.2021.

1. Bei der eingereichten Dissertation zu dem Thema “*Taking a closer look at the desirability of desirable difficulties—additionally focusing on required prerequisites, negative side-effects, and negative consequences*” handelt es sich um meine eigenständig erbrachte Leistung.

2. Anderer als der von mir angegebenen Quellen und Hilfsmittel habe ich mich nicht bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen veröffentlichten oder unveröffentlichten Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Dissertation oder Teile davon habe ich

☒ bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

☐ wie folgt an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt:

Titel der Arbeit:

Hochschule und Jahr:

Art der Prüfungs- oder Qualifikationsleistung:

Veröffentlicht in:

Es handelt sich dabei um folgenden Teil der Dissertation:

[Angabe in der Dissertation]

4. Die abgegebenen digitalen Versionen stimmen mit den abgegebenen schriftlichen Versionen überein.

5. Ich habe mich keiner unzulässigen Hilfe Dritter bedient und insbesondere die Hilfe einer kommerziellen Promotionsberatung nicht in Anspruch genommen.

6. Im Fall einer kumulativen Dissertation: Die Mitwirkung von Ko-Autor*innen habe ich durch eine von diesen unterschriebene Erklärung dokumentiert. Eine Übersicht, in der die einzelnen Beiträge nach Ko-Autor*innen und deren Anteil aufgeführt sind, füge ich anbei.

7. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

Gudensberg, 14.12.2021

Ort, Datum

M.Sc. Kristin Wenzel

Nummerierte Aufstellung der publizierten Artikel (Reihenfolge gemäß der Erwähnung in der Synopsis):

1. Reinhardt, N., Wenzel, K., & Reinhard, M.-A. (2019). Am I responsible for my learning success? A study about the correlation between locus of control and attitudes towards and self-reported use of desirable difficulties. *Journal of Psychological and Educational Research, 27*(1), 7–24.

2. Reinhard, M.-A., Weissgerber, S. C., & Wenzel, K.* (2019). Performance expectancies moderate the effectiveness of worked-examples and problem-solving over time. *Frontiers in Psychology, 10*, Article 1623. https://doi.org/10.3389/fpsyg.2019.01623 *shared first-authorship.

3. Wenzel, K., & Reinhard, M.-A. (2019). Relatively unintelligent individuals do not benefit from intentionally hindered learning: The role of desirable difficulties. *Intelligence, 77,* 101405. https://doi.org/10.1016/j.intell.2019.101405

4. Wenzel, K., & Reinhard, M.-A. (2021a). Does the end justify the means? Learning tests lead to more negative evaluations and to more stress experiences. *Learning and Motivation, 73*, Article 101706. https://doi.org/10.1016/j.lmot.2020.101706

5. Wenzel, K., & Reinhard, M.-A. (2020). Tests and academic cheating: do learning tasks influence cheating by way of negative evaluations?. *Social Psychology of Education, 23*(3), 721–753. https://doi.org/10.1007/s11218-020-09556-0

6. Wenzel, K., & Reinhard, M.-A. (2021b). Learning with a double-edged sword? Beneficial and detrimental effects of learning tests—taking a first look at linkages among tests, later learning outcomes, stress perceptions, and intelligence. *Frontiers in Psychology*, *12*, Article 693585. https://doi.org/10.3389/fpsyg.2021.693585

Darlegung des eigenen Beitrags an den publizierten Artikeln (im Verhältnis zu den Ko-Autor*innen; die weiteren Anteile/Beiträge verteilen sich jeweils auf den Ko-Autor bzw. auf die Ko-Autor*innen)**:**

Zu Nr. 1:

- Entwicklung der Konzeption: **in Teilen**
- Literaturrecherche: **in Teilen**
- Methodenentwicklung: **in Teilen**
- Entwicklung des Versuchsdesigns: **in Teilen**
- Datenerhebung: **in Teilen**
- Datenauswertung: **zu gleichen Teilen**
- Ergebnisdiskussion: **in Teilen**
- Erstellen des Manuskripts: **zu gleichen Teilen**
- Bewältigung des gesamten Review-Prozesses: **in Teilen**

Zu Nr. 2:

- Entwicklung der Konzeption: **kein Anteil**
- Literaturrecherche: **kein Anteil**
- Methodenentwicklung: **kein Anteil**
- Entwicklung des Versuchsdesigns: **kein Anteil**
- Datenerhebung: **kein Anteil**
- Datenauswertung: **kein Anteil**
- Ergebnisdiskussion: **in Teilen**
- Erstellen des Manuskripts: **in Teilen**
- Bewältigung des gesamten Review-Prozesses: **zu gleichen Teilen**

Zu Nr. 3:

- Entwicklung der Konzeption: **zu gleichen Teilen**
- Literaturrecherche: **vollständig**
- Methodenentwicklung: **überwiegend**
- Entwicklung des Versuchsdesigns: **mehrheitlich**
- Datenerhebung: **mehrheitlich**
- Datenauswertung: **überwiegend**
- Ergebnisdiskussion: **überwiegend**
- Erstellen des Manuskripts: **überwiegend**
- Bewältigung des gesamten Review-Prozesses: **überwiegend**

Zu Nr. 4:

- Entwicklung der Konzeption: **zu gleichen Teilen**
- Literaturrecherche: **vollständig**
- Methodenentwicklung: **überwiegend**
- Entwicklung des Versuchsdesigns: **zu gleichen Teilen**
- Datenerhebung: **überwiegend**
- Datenauswertung: **mehrheitlich**
- Ergebnisdiskussion: **überwiegend**
- Erstellen des Manuskripts: **überwiegend**
- Bewältigung des gesamten Review-Prozesses: **überwiegend**

Zu Nr. 5:

- Entwicklung der Konzeption: **überwiegend**
- Literaturrecherche: **vollständig**
- Methodenentwicklung: **vollständig**

- Entwicklung des Versuchsdesigns: **überwiegend**
- Datenerhebung: **überwiegend**
- Datenauswertung: **überwiegend**
- Ergebnisdiskussion: **überwiegend**
- Erstellen des Manuskripts: **überwiegend**
- Bewältigung des gesamten Review-Prozesses: **überwiegend**

Zu Nr. 6:

- Entwicklung der Konzeption: **überwiegend**
- Literaturrecherche: **vollständig**
- Methodenentwicklung: **vollständig**
- Entwicklung des Versuchsdesigns: **vollständig**
- Datenerhebung: **vollständig**
- Datenauswertung: **vollständig**
- Ergebnisdiskussion: **überwiegend**
- Erstellen des Manuskripts: **überwiegend**
- Bewältigung des gesamten Review-Prozesses: **überwiegend**

Anschriften der Ko-Autor*innen:

Prof. Dr. Marc-André Reinhard: reinhard@psychologie.uni-kassel.de

Dr. Sophia Weissgerber: scweissgerber@uni-kassel.de

Nina Reinhardt, M.Sc.: Nina.Reinhardt@uni-kassel.de

Unterschrift der Antragstellerin:

Gudensberg, 14.12.2021 _____

Ort, Datum M.Sc. Kristin Wenzel

Ich bestätige die von Frau Wenzel bezüglich ihres Eigenanteils abgegebenen Erklärungen:

1.

Prof. Dr. Marc-André Reinhard Unterschrift: …….

2.

Dr. Sophia Christin Weissgerber Unterschrift: …….

3.

Nina Reinhardt, M.Sc. Unterschrift: …….

Further Manuscripts and Publications

While I was working on this dissertation, I was also involved in further projects. Although the following four papers are not an integral part of my dissertation, they still deserve mention. Moreover, two of these papers were cited in my dissertation (Mariss et al., 2022; Wenzel et al., 2022).

Published articles (in peer-reviewed journals):

1. Wenzel, K., Schweppe, J., & Rummer, R. (2022*).* Are open-book tests still as effective as closed-book tests even after a delay of two weeks? *Applied Cognitive Psychology*, *36*(3), 699-707. https://doi.org/10.1002/acp.3943

2. Mariss, A.*, Wenzel, K.*, Grünberg, C., & Reinhard, M.-A. (2022). Who wants to learn harder? The relationship between conservatism and liberalism, desirable difficulties, and academic learning. *Social Psychology of Education, 25,* 209-248. https://doi.org/10.1007/s11218-021-09681-4, * shared first-authorship.

3. Schindler, S., Wenzel, K., Dobiosch, S., & Reinhard, M.-A. (2019). The role of belief in a just world for (dis)honest behavior. *Personality and Individual Differences*, *142*, 72–78. https://doi.org/10.1016/j.paid.2019.01.037

4. Wenzel, K., Schindler, S**.**, & Reinhard, M.-A. (2017). General belief in a just world is positively associated with dishonest behavior. *Frontiers in Psychology*, *8*, Article 1770. https://doi.org/10.3389/fpsyg.2017.01770

APPENDIX:

COPIES OF PUBLISHED ARTICLES

APPENDIX A

Reinhardt, N., Wenzel, K., & Reinhard, M.-A. (2019). Am I responsible for my learning success? A study about the correlation between locus of control and attitudes towards and self-reported use of desirable difficulties. *Journal of Psychological and Educational Research, 27*(1), 7–24.

This is the final article version published by Psychology Journals: JPER & IJEPC in *Journal of Psychological and Educational Research* available online: https://www.marianjournals.com/book/volume-27-issue-1-2019/

Journal of Psychological and
Educational Research
JPER - 2019, 27 (1), May, 7-24

AM I RESPONSIBLE FOR MY LEARNING SUCCESS? A STUDY ABOUT THE CORRELATION BETWEEN LOCUS OF CONTROL AND ATTITUDES TOWARDS AND SELF-REPORTED USE OF DESIRABLE DIFFICULTIES

Nina Reinhardt * **Kristin Wenzel** **Marc-André Reinhard**
University of Kassel, Germany

Abstract
Desirable difficulties are learning strategies that lead to more effective and durable learning even if the application produces difficulty at the moment of learning (Bjork, 1994). With this study we investigated the question which learner characteristics are linked to perception and application of those effortful strategies in self-regulated learning situations. In doing so we focused on locus of control, a construct that describes the extent to which individuals feel to have control over the outcome which arises from their own behavior (Rotter, 1975). Supporting our assumptions, internals-thus students with a stronger feeling of controllability - showed more positive attitudes towards desirable difficulties, whereas externals-thus students who suppose that external forces like chance/fate or powerful others have control over their behavior - showed no correlation. Furthermore, internals showed increased self-reported use of desirable difficulties. Contrary to prediction, external locus of control also was correlated to self-reported use of desirable difficulties. Results are discussed related to implications for the application of desirable difficulties in a real academic context regarding students with different learning characteristics, as well as to implications for further research.

Keywords: desirable difficulties; locus of control; learning; learner characteristics

Correspondence concerning this paper should be addressed to:

*Department of Psychology, University of Kassel. Address: Holländische Straße 36-38, 34127 Kassel, Germany. E-mail: *nina.reinhardt@uni-kassel.de*

Introduction

Concerning the application of effective learning strategies, individual differences or learner characteristics seem to be very important. While some students use relatively ineffective strategies other students use strategies which are mentally more challenging, but thereby promise a greater success of learning (*e.g.,* Bjork, Dunlosky, & Kornell, 2012; Karpicke, Butler, & Roediger, 2009). Such learning strategies are for instance so called *desirable difficulties.* The application of desirable difficulties leads to a more complex learning situation, but thereby especially benefits long-term effects regarding memory and transfer of the learned knowledge (Bjork, 1994). However, even if desirable difficulties promise greater learning success, most students do not use them in self-regulated learning situations. Instead, the majority prefers more simple strategies like for example repeated reading (Karpicke et al., 2009).

The present study wants to identify factors that are responsible for individual differences concerning the perception and application of desirable difficulties. Thereby, the main focus is on the *locus of control* construct, which is defined as the extent to which an individual perceives to have control over the outcome that arises from their own behavior (Rotter, 1975; Rotter, Chance, & Phares, 1972). Given that the application of desirable difficulties leads to more challenging learning situations - because they elicit more time and cognitive effort - locus of control can be expected as a learner characteristic that promotes the feeling of being able to cope those complex situations because the individual feels confident that his or her learning effort will result in greater learning success (*e.g.*, Prociuk & Breen, 1997).

Desirable difficulties

Desirable difficulties are learning strategies that lead to more effective and durable learning through producing difficulty at the moment of application (Bjork, 1994). Recent research indicated that different forms of desirable difficulties exist, often termed as *spacing*, *interleaving*, *generation* and *testing*. Spacing or distributed learning means to separate specific learning sessions in different units in contrast to learning one topic from beginning to end (*e.g.*, Vlach & Sandhofer, 2012). In some way interleaving contains this strategy but additionally includes that in between the temporary distributed learning units different topics get mixed (*e.g.*, Ziegler & Stern, 2014). For visualizing the

difference, one can imagine that spacing or distributed learning means first learning unit AA, then BB and at least CC, whereas interleaving means learning first A, B, C and then again A, B, C. In particular, both strategies promote long-term memory (*e.g.*, Vlach & Sandhofer, 2012; Ziegler & Stern, 2014). Generation further describes active generation of learning materials and predictions instead of merely (re-)reading and repeating the material which also strengthens long-term memory (*e.g.*, Bertsch, Pesta, Wiscott, & McDaniel, 2007; DeWinstanley & Bjork, 2004; Slamecka & Graf, 1978). Testing requires active memory recall and retrieval rather than simply rehearsing the relevant knowledge (*e.g.*, Karpicke et al., 2009; for recent meta-analyses *see*: Adesope, Trevisan, & Sundararajan, 2017; Rowland, 2014).

While positive effects of desirable difficulties on learning success are relatively well understood and seen as rather robust, more research regarding personality variables and learner characteristics that may affect the perception and application of desirable difficulties is still valuable (*e.g.*, Weissgerber, Reinhard, & Schindler, 2016).

Internal and external locus of control

Locus of control describes the individual tendency to attribute causes of life events to one's own behavior or external influences (Rotter, 1975). The origin of this construct lies in *social learning theory* (Rotter et al., 1972). According to this theory there are two extremes - an internal and an external locus of control (scores in between these extremes can of course also exist). Individuals with a more internal locus of control assume a strong connection between their behavior and the thereby resulting outcome. They are often dubbed as *internals*. Individuals with a more external locus of control instead suppose that external forces like chance/fate or powerful others have control over their lives. They are often dubbed as *externals* (Rotter, 1975). Depending on the conceptualization the construct can be measured two-dimensional (*e.g.*, Rotter, 1975) or three-dimensional (*e.g.*, Levenson, 1973) by dividing the external extreme in two different dimensions (i.e. external locus of control because of powerful others or external locus of control because of chance). Rotter further defined locus of control as a *generalized expectancy* which only affects human behavior in complex and ambiguous situations in comparison to specific expectancies which lead individuals' trough familiar situations (Rotter, 1975).

In the past, research has often associated locus of control with specific performance situations (Spector, 1982). Several experiments have shown that internals outperform externals in actively seeking relevant information and in the utilization of those information (*e.g.*, Phares, 1968; Ude & Vogler, 1969; Wolk & DuCette, 1973). Based on this, internals are generally considered to be more successful and effective. Numerous experiments confirmed this assumption by operationalizing professional success for instance in form of salary level or job position (*e.g.*, Valecha, 1972) as well as by judgements of supervisors (*e.g.*, Majumder, MacDonald, & Greever, 1977). Better information seeking and processing behavior as well as greater success (better grades) were also linked with higher internal locus of control in academic context (Prociuk & Breen, 1997). However, within the research body of desirable difficulties - to our knowledge - the construct of locus of control has never been considered before.

The current research

In the following we want to argue why locus of control could be an important factor that may influence individuals' perception and application of desirable difficulties: Given that the application of desirable difficulties lead to cognitively effortful learning processes (Bjork et al., 2012), then, especially in those complex, ambiguous, and relatively unknown situations generalized expectancies like locus of control should determine human behavior (Rotter, 1975). Given that individuals with a more internal locus of control generally act more effective and successful (*e.g.*, Majumder et al., 1977; Prociuk & Breen, 1997; Valecha, 1972), and additionally have better skills in seeking or utilizing relevant information (*e.g.*, Phares, 1968; Ude & Vogler, 1969; Wolk & DuCette, 1973), it can be assumed that for internals the application of desirable difficulties isn't as (mentally) challenging as for externals. Thus, an internal locus of control could facilitate affective and cognitive perception regarding the application of desirable difficulties. Regarding the external dimension, no correlation to attitudes towards desirable difficulties is assumed. Externals don't believe in the effectiveness of their behavior. Therefore, to them it should make no difference which specific learning strategy they use in order to reach their learning goal. In their opinion they are not able to reach it by their own efforts because they perceive it depending on powerful others, luck, chance and/or fate.

Therefore, we first predict that students who score higher in internal locus of control have more positive attitudes towards desirable difficulties (Hypothesis

1a), whereas students who score higher in external locus of control (because of powerful others or chance) show no correlation to attitudes towards desirable difficulties (Hypothesis 1b). Given that assumption, this should implicate an increased application of desirable difficulties only for internals but not for externals. Therefore, we second predict that students who score higher in internal locus of control show increased self-reported use of desirable difficulties (Hypothesis 2a), whereas students who score higher in external locus of control (because of powerful others or chance) show no correlation to self-reported use of desirable difficulties (Hypothesis 2b).

Methods

Participants

In total 504 participants (54% male) completed our online-survey. Participants mean age was 26.58 (SD=6.83) and ranged from 16 to 68 years. The majority (96%) of the participants were inhabitants of the USA. Participation requirements contained that all subjects were registered students at the moment of questioning. The recruitment took place via Amazon Mechanical Turk.

Measures

All of the measures were administered in English. The IPC-Scale developed by Levenson (1973) was used to measure *locus of control*. Levenson conceptualizes locus of control as a three-dimensional construct by dividing the external extreme in two different categories. The I-Scale serves to measure the internal extreme (α=.788; *e.g.*, „My life is determined by my own actions“). The P-Scale measures perception of external control which arises from individuals feelings that powerful others control their life events (α=.880; *e.g.*, „Getting what I want requires pleasing those people above me“). The C-Scale measures the external fatalistic control which is attributed with luck, fate and chance (α=.878; *e.g.*, „Whether or not I get into a car accident is mostly a matter of luck“). Participants were instructed to rate each statement on a 6-point Likert response scale (1=*strongly disagree* and 6=*strongly agree*).

Dispositional stress as a control variable was assessed by Cohen's and Williams's 10-item Perceived Stress Scale (1988). The questionnaire measures the degree to which specific life events and situations are experienced as stressful. General statements like for example "how often have you felt that

things were going your way" or "how often have you felt nervous and stressed" (α=.859) should be assessed on a 5-point Likert response scale (1=*never* and 5=*very often*) regarding their probability of occurrence during the last month. The measure of stress serves as a control variable to check whether the attitude and the decision to use/not use mentally challenging learning strategies is dependent of a specific level of stress which is common for students in their academic environment (*e.g.*, Abouserie, 1994; Anda et al., 2000; Heins, Fahey, & Leiden, 1984).

S*elf-efficiency beliefs* as a control variable were assessed using the General Self-Efficacy Scale by Schwarzer and Jerusalem (1995). The scale measures the extent to which someone beliefs to be able to handle critical situations by oneself. Ten items (α=.866; *e. g.*, „I can usually handle whatever comes my way") should be rated on a 4-point Likert response scale (1=*not at all true* and 4=*exactly true*).

Attitudes towards and self-reported use of desirable difficulties was assessed with items used by Weissgerber et al. (2016; *see also* Weissgerber, Reinhard, & Schindler, 2018). The scale concentrates on five different desirable difficulties: self-generation of learning contents, labeled as *generation*, generation of predictions, labeled as *predictions*, *self-testing*, *spacing/ interleaving* and *practicing*, which means a mixed form of self-testing with self-generated learning materials. Each of these five different desirable difficulties types is captured with three items. Two items assessed the attitude towards each type of desirable difficulty (*e.g.*, „I like to create my own learning materials"; „I think it is useful to acquire knowledge myself"). Self-reported use in self-regulated learning situations was assessed for each type of desirable difficulty with one item (*e.g.*, „Compared to other learning methods, I work out the subject matter myself"). Participants were instructed to rate each statement on a 7-point Likert response scale (1=*totally disagree* and 7=*totally agree*). Both, the attitudes scale (α=.890) and the self-reported use scale (α=.805) showed good reliability.

Procedure

The survey contained three different parts. After participants gave their permission to participate voluntary and reported that they were registered college or university students, we assessed their demographics (*i.e.*, gender, age, homeland). The second part included the assessment of locus of control, dispositional stress and self-efficiency beliefs in randomized order. During the

third part, the measure of attitudes toward and self-reported use of desirable difficulties took place. At the end of the online survey the respondents were thanked for their participation and got $0.60 as a reward.

Results

Table 1 shows the means, standard deviations and correlation coefficients among all variables examined in this study. All correlations coefficients were Bonferroni adjusted. The total value for attitudes exhibited a mean value of 5.10 (SD=1.01) and for self-reported use a mean value of 5.05 (SD=1.06). Mean value of internal locus of control was 4.32 (SD=0.70), mean value of external locus of control which arises from the feeling that powerful others hold control above life events was 3.60 (SD=1.02) and mean value of external fatalistic locus control was 3.50 (SD=1.02). Stress showed a mean value of 2.82 (SD=0.70) and self-efficacy beliefs a mean value of 3.09 (SD=0.51).

Internal locus of control correlated significantly with attitudes towards desirable difficulties ($r=.549$, $p<.001$) and with self-reported use of desirable difficulties ($r=.526$, $p<.001$). Thus, internal locus of control is related to positive attitudes towards and increased self-reported use of desirable difficulties. Neither external locus of control because of powerful other, nor the external fatalistic dimension was found to be correlated with attitudes towards desirable difficulties and self-reported use of desirable difficulties. Both external locus of control dimensions also showed no correlation to internal locus of control.

Regarding measured control variables both showed significant correlations with the dependent variables. Stress correlated significantly with attitudes towards ($r=-.220$, $p<.001$) and self-reported use of desirable difficulties ($r=-.193$, $p<.001$), both negatively. Thus, lower levels of stress are related to positive attitudes towards and increased self-reported use of desirable difficulties. Self-efficiency beliefs also showed significant correlations with attitudes towards desirable difficulties ($r=.498$, $p<.001$) and with self-reported use of desirable difficulties ($r=.477$, $p<.001$). Thus, higher values of self-efficiency beliefs are related to positive attitudes towards and increased self-reported use of desirable difficulties. Furthermore, there was a negative correlation between stress and self-efficiency beliefs ($r=-.537$, $p<.001$).

Table 1. Intercorrelations, Means, Standard Deviations and Cronbachs α for Attitudes Toward and Self-Reported Use of Desirable Difficulties, Locus of Control, Stress, and Self-Efficiency Beliefs (N=504)

	1.	2.	3.	4.	5.	6.	7.	M	SD	α
1. Attitudes	1							5.10	1.01	.890
2. Self-reported use	.884**	1						5.05	1.06	.805
3. Internal	.549**	.526**	1					4.32	0.70	.788
4. Powerful others	.051	.096	.095	1				3.60	1.02	.880
5. Chance	.046	.103	.048	.838**	1			3.50	1.02	.878
6. Stress	-.220**	-.193**	-.298**	.474**	.513**	1		2.82	0.70	.859
7. Self-efficiency	.498**	.477**	.581**	-.210**	-.216**	-.537**	1	3.09	0.51	.866

Note: *p<.05 (two-tailed); **p<.001 (two-tailed); all calculated correlations were Bonferroni adjusted; Attitudes=attitudes towards desirable difficulties; Self-reported use=Self-reported use of desirable difficulties; Internal=Internal locus of control dimension; Powerful others=External locus of control dimension because of powerful others; Chance=External locus of control because of chance; Self-efficiency=Self-efficiency beliefs.

Note that a more detailed look to the different kinds of desirable difficulties reveals quiet the same correlations for the specific scores (see Appendix Table A). Correlations with internal locus of control all reached significance and showed a positive direction (r=.566, p<.001 to r=.705, p<.001). Correlations with external locus of control predominantly reached no significance with specific scores of desirable difficulties. Only interleaving / spacing showed a significant correlation with external fatalistic locus of control (r=.161, p<.05).

To test Hypothesis 1a which claimed that students who score higher in internal locus of control have more positive attitudes towards desirable difficulties and Hypothesis 1b which claimed that students who score higher in external locus of control (because of powerful others or chance) show no correlation to attitudes towards desirable difficulties, already the presented correlations indicated support (Table 1). Thus, internal locus of control was positively linked to attitudes towards desirable difficulties (r=.549, p<.001) but neither the external dimension because of powerful others, nor the external fatalistic dimension exhibited significant correlations. To strengthen these findings, a multiple, stepwise regression analysis was conducted. Respectively, attitudes towards desirable difficulties was used as criterium and the three dimensions of locus of control were used as predictors. Additionally, in a second and third step, gender, age, self-efficiency beliefs and stress were added as control variables to test if the effect of internal locus of control would remain robust (Table 2).

Table 2. Multiple, Stepwise Regression Analysis with Attitudes Toward Desirable Difficulties as the Criterion and Locus of Control, Gender, Age, Self-Efficiency Beliefs, and Stress as Predictors (N=504)

		Parameter estimates						
Model	Predictors	*B*	*SE*	β	*t*	*p*	R^2	$r_{x(y, z)}$
(1)	Internal	0.821	0.056	.551	14.673	<.001	.303	.548
	Powerful others	-0.060	0.068	-.061	-0.880	.379		-.033
	Chance	0.070	0.068	.070	1.026	.305		.038
(2)	Internal	0.828	0.056	.556	14.712	<.001	.307	.549
	Powerful others	-0.047	0.068	-.047	-0.686	.493		-.026
	Chance	-0.068	0.068	.069	1.011	.312		.038
	Age	-0.003	0.006	-.019	-0.493	.623		-.018
	Gender	-0.131	0.078	-.065	-1.690	.092		-.063
(3)	Internal	0.553	0.069	.371	8.028	<.001	.362	.288
	Powerful others	0.015	0.067	.016	0.232	.817		.008
	Chance	0.090	0.067	.091	1.345	.179		.048
	Age	-0.003	0.005	-.020	-0.541	.589		-.019
	Gender	-0.133	0.075	-.066	-1.773	.077		-.064
	Self-efficiency	0.603	0.099	.306	6.088	<.001		.218
	Stress	0.001	0.071	.000	0.009	.992		.000

Note: $r_{x(y, z)}$=Semipartial correlation coefficients; Internal=Internal locus of control dimension; Powerful others=External locus of control dimension because of powerful others; Chance=External locus of control because of chance; Self-efficiency=Self-efficiency beliefs.

As shown in Table 2 (model 1) internal locus of control significantly predicted attitudes towards desirable difficulties (β=.551, p<.001). The model explained 30.3% of variance, F(3, 500)=72.394, p<.001, R^2=.303, $R^2_{adjusted}$=.299. Gender- or age-effects did not exist. Thus, neither gender nor age did change the effects of internal locus of control as a predictor (model 2). However, it should be pointed out that self-efficiency beliefs (model 3) also had a significant impact on attitudes towards desirable difficulties (β=.306, p<.001). In the shared model (model 3) the impact of internal locus of control decreased but still reached a significant level (β=.371, p<.001). The shared model explained 36.2% of variance, F(7, 496)=40.219, p<.001, R^2=.362, $R^2_{adjusted}$=.353. In all of the three models neither external locus of control because of powerful others, nor external fatalistic locus of control showed a significant effect to attitudes towards desirable difficulties.

To test Hypothesis 2a which claimed that students who score higher in internal locus of control show increased self-reported use of desirable difficulties and Hypothesis 2b which claimed that students who score higher in external locus of control (because of powerful others or chance) show no correlation to self-reported use of desirable difficulties, again, the already presented

correlations indicated support (Table 1). Thus, internal locus of control was positively linked to attitudes towards desirable difficulties (r=.549, p<.001) but neither the external dimension because of powerful others, nor the external fatalistic dimension exhibited significant correlations. For further analysis a multiple, stepwise regression analysis was conducted. Respectively, self-reported use of desirable difficulties was used as criterium and the three dimensions of locus of control were used as predictors. Additionally, in a second and third step, gender, age, self-efficiency beliefs and stress were added as control variables to test if the effect of internal locus of control would remain robust (Table 3).

Table 3. Multiple, Stepwise Regression Analysis with Self-Reported Use of Desirable Difficulties as the Criterion and Locus of Control, Gender, Age, Self-Efficiency Beliefs, and Stress as Predictors (N=504)

		Parameter estimates						
Model	Predictors	*B*	*SE*	β	*t*	*p*	R^2	$r_{x(y,\,z)}$
(1)	Internal	0.826	0.060	.526	13.801	<.001	.284	.522
	Powerful others	-0.063	0.073	-.061	-0.871	.384		-.033
	Chance	0.134	0.072	.128	1.844	.066		.070
(2)	Internal	0.834	0.060	.531	13.836	<.001	.286	.524
	Powerful others	-0.052	0.073	-.050	-0.710	.478		-.027
	Chance	0.133	0.072	.127	1.832	.068		.069
	Age	0.000	0.006	.002	0.062	.950		.002
	Gender	-0.110	0.083	-.052	-1.328	.185		-.050
(3)	Internal	0.532	0.074	.338	7.215	<.001	.345	.262
	Powerful others	0.018	0.071	.017	0.255	.799		.009
	Chance	0.161	0.072	.154	2.248	.025		.082
	Age	0.000	0.006	.001	0.034	.973		.001
	Gender	-0.114	0.080	-.054	-1.425	.155		-.052
	Self-efficiency	0.648	0.106	.312	6.123	<.001		.222
	Stress	-0.020	0.076	-.013	-0.260	.795		-.009

Note: $r_{x(y,\,z)}$=Semipartial correlation coefficients; Internal=Internal locus of control dimension; Powerful others=External locus of control dimension because of powerful others; Chance=External locus of control because of chance; Self-efficiency=Self-efficiency beliefs.

As shown in Table 3 (model 1) internal locus of control significantly predicted self-reported use of desirable difficulties (β=.526, p<.001). The model explained 28.4% of variance, F(3, 500)=66.029, p<.001, R^2=.284, $R^2_{adjusted}$=.279. Gender- or age-effects did not exist. Thus, whether gender nor age did change the effects of internal locus of control as a predictor (model 2). But it should be pointed out that self-efficiency beliefs (model 3) also had a significant impact on

self-reported use of desirable difficulties (β=.312, p<.001) as well as the external fatalistic dimension of locus of control labeled as chance (β=.154, p<.025). In the shared model (model 3) the impact of internal locus of control decreased but still reached a significant level (β=.328, p<.001). The shared model explained 34.5% of variance, F(7, 496)=28.115, p<.001, R^2=.345, $R^2_{adjusted}$=.336.

Hypothesis 1a claimed that students who score higher in internal locus of control have more positive attitudes towards desirable difficulties and Hypothesis 1b claimed that students who score higher in external locus of control (because of powerful others or chance) show no correlation to attitudes towards desirable difficulties. Regression analysis revealed a significant impact of internal locus of control on attitudes towards desirable difficulties - note that the effects remain robust although adding control variables (*i.e.*, gender, age, self-efficiency beliefs, and stress) - and no significant effects of both external dimensions of locus of control (Table 2). Therefore, Hypothesis 1a and 1b can be accepted. Hypothesis 2a claimed that students who score higher in internal locus of control show increased self-reported use of desirable difficulties and Hypothesis 2b claimed that students who score higher in external locus of control (because of powerful others or chance) show no correlation to self-reported use of desirable difficulties. Regression analysis revealed a significant impact of internal locus of control on self-reported use of desirable difficulties - note that the effects remain robust although adding control variables (*i.e.*, gender, age, self-efficiency beliefs, and stress) - but also a significant effect of external fatalistic locus of control (Table 3). Therefore, only Hypothesis 2a can be accepted whereas Hypothesis 2b must be rejected.

Conclusions

This study was initiated to examine the relationship between perceptions of desirable difficulties and personality factors, respectively learner characteristics. Our research project seemed to be necessary because although positive effects of desirable difficulties on learning success are relatively well understood through past research, personality variables that may affect perceptions and applications of such desirable difficulties still need further exploration (*e.g.*, Weissgerber et al., 2018). The present study focused on locus of control, a personality construct that describes the extent to which individuals feel to have control over the outcome that arises from their own behavior (Rotter,

1975). We assumed a more internal locus of control to be linked to positive attitudes towards and increased self-reported use of desirable difficulties, whereas a more external locus of control should not be linked to attitudes towards and self-reported use of desirable difficulties.

The reported findings could support our first three assumptions: Our results provided significant correlation between internal locus of control and attitudes towards desirable difficulties (Hypothesis 1a), as well as no correlation between external locus of control (because of powerful others or chance) and attitudes towards desirable difficulties (Hypothesis 1b). There was also a significant correlation between internal locus of control and self-reported use of desirable difficulties (Hypothesis 2a). These findings are in line with past research, which mostly demonstrated that individuals with a more internal locus of control generally act more effective and successful (*e.g.*, Majumder et al., 1977; Prociuk & Breen, 1997; Valecha, 1972) and have better skills in seeking or utilizing relevant information (*e.g.*, Phares, 1968; Ude & Vogler, 1969; Wolk & DuCette, 1973). Thus, we assumed that for internals the application of effortful learning strategies isn't as (mentally) challenging as for externals, which in turn could facilitate affective and cognitive perception of such difficulties only for internals but not for externals. Presented correlations (Table 1) and multiple, stepwise regression analyses (Table 2 and Table 3) confirmed these assumptions. Newly - and surely the added value of this study - is the transferability of the results on the learning context respectively the context of attitudes towards and self-reported use of desirable difficulties.

Contrary to our assumptions, the presented regression analysis (Table 3) showed external fatalistic locus of control to be a significant predictor of self-reported use of desirable difficulties. We did not predict this because we theoretically assumed students with a more external locus of control to not believe in the effectiveness of their own behavior, so that they should not show preferences for specific learning strategies. Still, the external fatalistic dimension showed a significant correlation to self-reported use of desirable difficulties so that we cannot support Hypothesis 2b.

Taking a more detailed look at the correlations between external fatalistic locus of control and the specific desirable difficulties, only *interleaving/spacing* reached significance, whereas the other kinds (in line with our expectation) were not correlated with external fatalistic locus of control (*see* Appendix Table A). In comparison to the other specific kinds of desirable difficulties,

interleaving/spacing requires less activity than the other ones because it only consists of mixing different topics throughout different learning sessions. Because less proactivity is needed, maybe this strategy isn't that challenging as the other ones and therefore fatalistic externals believe that they can handle it with luck, fate and chance. Therefore, our hypothesis possibly cannot be applied to this specific kind of desirable difficulties because interleaving/spacing isn't seen as a real difficulty. Further, we advise to be careful by interpreting external fatalistic locus of control as predictor for self-reported-use of desirable difficulties because the external fatalistic dimension only reached significance in a shared/larger model (model 3) within which the control variables stress and self-efficiency beliefs were integrated (Table 3). Thus, it seems that the external fatalistic dimension shares some variance with stress and/or self-efficiency beliefs. Individuals scoring higher on the external fatalistic dimension could possibly be less sensitive to stress, insofar as that they believe their success to be dependent of fatalistic causes like chance and not on their own abilities. Future studies focusing on the relation between external locus of control and stress would be valuable to further explore this linkage.

In general, we want to highlight that internal locus of control significantly correlated with all kinds of desirable difficulties and that regression analyses revealed that the major impact on self-reported use of desirable difficulties emanates from internal rather than external fatalistic locus of control. Due to this, we consider internal locus of control as the more important factor regarding affective and cognitive perception of desirable difficulties. Hence, the following practical implications will only refer to the role of internal locus of control.

Altogether, the findings are relevant for students in self-determined learning situations as well as for teachers whose aim it should be to promote the long-term learning success of their students. Teachers should be aware that different students, with their different learner characteristics, show different learning styles and that students (external) locus of control could be one important factor why some of them use more ineffective strategies, while others make greater use of more effective strategies. Given that internals and externals benefit equal from the application of desirable difficulties, especially insecure students which may feel overstrained (because of their external locus of control) should be sensitized above the superior learning benefits of desirable difficulties. Besides, teacher could help externals as well as internals using desirable

difficulties for instance through presenting them more opportunities to do so in the school or university context.

However, one limitation of our study is that we focused on correlations and not on causal effects. Thus, future studies should try to test our results using manipulations and not only observing relations. In line with this, even if our study showed that individuals with a more internal locus of control had more positive perceptions and increased applications of those strategies, it did not automatically evidence that internals also achieve greater learning success per se. The effect of locus of control on learning success due to the application of desirable difficulties has not yet been clarified. Thus, it seems to be interesting for further research to examine if locus of control acts as moderator in the relation between application of desirable difficulties and learning success in a real academic context. If internals and externals for instance don't benefit equally from the application of desirable difficulties - for example because internals show better skills in seeking relevant information (*e.g.*, Phares, 1968; Ude & Vogler, 1969; Wolk & DuCette, 1973) - a new question follows: In this case it seems to be valuable that teachers could be able to shift locus of control of their students to a more internal direction to promote their learning success. Thus, further research should examine if locus of control is changeable in an academic context.

Furthermore, it seems to be important to focus on the difference between locus of control and self-efficiency beliefs. In both multiple regression analyses self-efficiency beliefs showed significant impact on attitudes towards desirable difficulties (however, the effects of internal locus of control remained significant even while controlling for self-efficiency beliefs). Indeed, self-efficiency belief coefficients were even relatively smaller than coefficients of internal locus of control. Hence, further research should focuse on the difference between locus of control and self-efficiency beliefs.

Funding

This research was supported by a LOEWE grant from the Hessian Ministry for Science and the Arts entitled "desirable difficulties; intrinsic cognitive motivation and performance expectancies" awarded to Marc-André Reinhard.

References

Abouserie, R. (1994). Sources and levels of stress in relation to locus of control and self-esteem in university students. *Educational Psychology, 14*, 323-330. doi: 10.1080/0144341940140306

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*, 659-701. doi: 10.3102/0034654316689306

Anda, D., Baroni, S., Boskin, L., Buchwald, L., Morgan, J., Ow, J., … Weiss, R. (2000). Stress, stressors and coping among high school students. *Children and Youth Services Review, 22*, 441-463. doi: 10.1016/S0190-7409(00)00096-7

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*, 201-210. doi: 10.3758/BF03193441

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge: MIT Press.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2012). Self-regulated learning: beliefs, techniques, and illusions. *The Annual Review of Psychology*, *64*, 417-445. doi: 10.1146/annurev-psych-113011-143823

Cohen, S., & Williams, G. M. (1988). Perceived stress in a probability sample in the United States. In S. Spacapan, & S. Oskamp (Eds.). *The social psychology of health* (pp. 31-67). Newbury Park: Sage.

DeWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition, 32*, 945-955. doi: 10.3758/BF03196872

Heins, M., Fahey, S. N., & Leiden, L. I. (1984). Perceived stress in medical, law, and graduate students. *Journal of Medical Education, 59*, 169-179. doi: 10.1016/0277-9536(89)90351-1

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17*, 471-479. doi: 10.1080/09658210802647009

Levenson, H. (1973). Activism and powerful others: Distinctions within the concept of internal-external control, *Journal of Personality Assessment, 38*, 377-383. doi: 10.1080/00223891.1974.10119988

Majumder, R. K., MacDonald, A. P., & Greever, K. B. (1977). A study of rehabilitation counselors locus of control and attitudes toward the poor. *Journal of Counseling Psychology*, *24,* 137-141. doi: 10.1037/0022-0167.24.2.137

Phares, E. J. (1968) Differential utilization of information as a function of internal-external control. *Journal of Personality*, *36*, 649-661. doi: 10.1111/j.1467-6494.1968.tb01498.x

Prociuk, T. J., & Breen, L. J. (1997). Internal-external locus of control and information-seeking in a college academic situation. *The Journal of Social Psychology, 101,* 309-310. doi: 10.1080/00224545.1977.9924022

Rotter, J. B. (1975). Some problems and misconceptions related to the construct of internal versus external control of reinforcement. *Journal of Consulting and Clinical Psychology, 43*, 56-67. doi: 10.1037/h0076301

Rotter, J. B., Chance, J. E., & Phares, E. J. (1972). *Applications of a social learning theory of personality.* New York: Holt, Rinehart and Winston, Inc.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*, 1432-1463. doi: 10.1037/a0037559

Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. In J. Weinman, S. Wright, & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35-37). Windsor: NFER-Nelson.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592-604. doi: 10.1037/0278-7393.4.6.592

Spector, P. E. (1982). Behavior in organizations as a function of employee's locus of control. *Psychological Bulletin, 91*, 482-497. doi: 10.1037/0033-2909.91.3.482

Ude, L. K., & Vogler, R. E. (1969). Internal versus external control of reinforcement and aware ness in a conditioning task. *The Journal of* Psychology, *73*, 63-67. doi: 10.1080/00223980.1969.10543517

Valecha, G. K. (1972). Construct validation of internal-external locus of reinforcement related to work-related variables. *Proceedings of the Annual Convention of the American Psychological Association, 7,* 455-456.

Vlach, H. A., & Sandhofer, C. M. (2012). Distributing learning over time: The spacing effect in children's acquisition and generalization of science

concepts. *Child Development, 83*, 1137-1144. doi: 10.1111/j.1467-8624.2012.01781.x

Weissgerber, S. C., Reinhard, M.-A., & Schindler, S. (2016). Study harder? The relationship of achievement goals to attitudes and self-reported use of desirable difficulties in self-regulated learning. *Journal of Psychological and Educational Research, 24*, 42-60.

Weissgerber, S. C., Reinhard, M.-A., & Schindler, S. (2018). Learning the hard way: Need for cognition influences attitudes toward and self-reported use of desirable difficulties. *Educational Psychology, 38,* 176-202. doi: 10.1080/01443410.2017.1387644

Wolk, S., & DuCette, J. (1973). The moderating effect of locus of control in relation to achievement-motivation variables. *Journal of Personality, 41,* 59-70. doi: 10.1111/j.1467-6494.1973.tb00660.x

Ziegler, E., & Stern, E. (2014). Delayed benefits of learning elementary algebraic transformations through contrasted comparisons. *Learning and Instruction*, *33,* 131-146. doi: 10.1016/j.learninstruc.2014.04.006

Received December 4, 2018
Revision February 28, 2019
Accepted May 22, 2019

Appendix

Intercorrelations, Means, Standard Deviations and Cronbachs α for Single Desirable Difficulties, Attitudes, Self-Reported Use, Locus of Control, Stress, and Self-Efficiency Beliefs (N = 504)

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	M	SD	α
1.	1.												5.11	1.16	.740
2.	.705**	1.											5.10	1.16	.780
3.	.643**	.706**	1.										5.10	1.24	.825
4.	.566*	.614**	.630**	1.									5.30	1.20	.829
5.	.588**	.606**	.649**	.663**	1.								4.85	1.17	.734
6.	.826**	.855**	.858**	.807**	.806**	1.							5.10	1.01	.890
7.	.770**	.804**	.810**	.798**	.816**	.884**	1.						5.05	1.06	.805
8.	.430**	.480**	.481**	.479**	.470**	.549**	.526**	1.					4.32	0.70	.788
9.	.064	.081	.001	.001	.146	.051	.096	.095	1.				3.60	1.02	.880
10.	.050	.076	.008	-.006	.161*	.046	.103	.048	.838**	1.			3.50	1.02	.878
11.	-.125	-.193**	-.257**	-.185**	-.147	-.220**	-.193**	-.298**	.474**	.513**	1.		2.82	0.70	.859
12.	.375**	.463**	.468**	.438**	.375**	.498**	.477**	.581**	-.210**	-.216**	-.537**	1.	3.09	0.51	.866

Note. *$p < .05$ (two-tailed); **$p < .001$ (two-tailed); all calculated correlations were Bonferroni adjusted; 1. = Generation, 2. = Prediction; 3. = Self-testing; 4. = Practicing; 5. = Interleaving/Spacing; 6. = attitudes towards desirable difficulties; 7. = Self-reported use of desirable difficulties; 8. = Internal locus of control dimension; 9. = External locus of control dimension because of powerful others; 10. = External locus of control because of chance; 11. = Stress; 12. = Self-efficiency beliefs.

APPENDIX B

Reinhard, M.-A., Weissgerber, S. C., & Wenzel, K.* (2019). Performance expectancies moderate the effectiveness of worked-examples and problem-solving over time. *Frontiers in Psychology, 10*, Article 1623. https://doi.org/10.3389/fpsyg.2019.01623 *shared first-authorship.

This is the final article version published by Frontiers in *Frontiers in Psychology* available online: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01623/full

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 August 2019
doi: 10.3389/fpsyg.2019.01623

Performance Expectancies Moderate the Effectiveness of More or Less Generative Activities Over Time

Marc-André Reinhard†, Sophia Christin Weissgerber† and Kristin Wenzel†*

Department of Psychology, University of Kassel, Kassel, Germany

We examined if the benefits of generation for long-term learning depend on individual differences in performance expectancies (PEs) prior to learning. We predicted that a greater generative activity (problem-solving) compared to less generative activity (worked-examples) should be more effective for pupils with higher PEs, especially in the long run. As a comparison group for problem-solving, we implemented a special type of worked-examples that decreased engaging in self-explanations, because our main prediction focused on PEs moderating the long-term effectivity of less versus greater generative activities. We tested students' immediate and delayed performance (after 3 months) using coherent curricular materials on linear functions in a sample of eighth graders (advanced school track). The results were partly in line with our predictions: Although we found no moderation of PE and generative activity, we obtained the predicted 3-way interaction of PE, generative activity, and time. Immediately, greater generative activity (problem-solving) was beneficial for pupils with higher PEs, while for pupils with lower PEs, problem-solving versus worked-examples did not differ. In the delayed test, this pattern reversed: for lower PEs, greater generative activity outperformed less generative activities, but there was no difference for higher PEs. Unexpectedly, the initial advantage of problem-solving for higher PEs could not be maintained, decreasing over three subsequent months, whereas the performance in the worked-example condition remained at a comparable level for higher PEs. The change in performance in the problem-solving condition for lower PEs was descriptively less pronounced than in the worked-example condition, but statistically not different. We further investigated the effects of problem-solving and worked-examples on changes in PEs after learning and after testing, hinting at gradual decrease in PEs and greater metacognitive accuracy in the problem-solving condition due to a reduction of overconfidence.

Keywords: desirable difficulties, problem-solving, generation effect, worked-examples, performance expectancies, meta-cognition, long-term learning

OPEN ACCESS

Edited by:
Huib Tabbers,
Erasmus University Rotterdam, Netherlands

Reviewed by:
Julian Roelle,
Ruhr University Bochum, Germany
Ouhao Chen,
Nanyang Technological University, Singapore
Martine Baars,
Erasmus University Rotterdam, Netherlands

***Correspondence:**
Marc-André Reinhard
reinhard@psychologie.uni-kassel.de

†Shared first authorship

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *28 June 2018*
Accepted: *27 June 2019*
Published: *21 August 2019*

Citation:
Reinhard M-A, Weissgerber SC and Wenzel K (2019) Performance Expectancies Moderate the Effectiveness of More or Less Generative Activities Over Time. Front. Psychol. 10:1623. doi: 10.3389/fpsyg.2019.01623

INTRODUCTION

The idea to trouble a learner by a difficult learning task may appear strange. Intuitively, wouldn't one ease the learning task to match the learner's achievement prediction in hope of raising said learner's achievement prospects? Yet, a growing body of research on a phenomenon dubbed "desirable difficulties" (e.g., Bjork, 1994; Bjork and Bjork, 2011) indeed supports such a seemingly

odd learning approach. The label "desirable difficulties" subsumes various learning conditions which require considerable but manageable effort to foster long-term learning. Although the introduced difficulties may not be beneficial for the short term, overcoming the difficulties may induce desirable cognitive processes and strengthen memory, thus paying off in the long run (e.g., Bjork and Bjork, 1992, 2011; Bjork, 1994).

It is often theorized that such learning gains can be attributed to stimulations of cognitive processes that increase an understanding and deeper encoding of information, and that desirable difficulties anchor information in long-term memory (e.g., Bjork and Bjork, 1992; Bjork, 1994). The kind of processing required of a difficult learning task and the processing used by the learner are identified as two central aspects regarding the desirability of a difficulty (McDaniel and Butler, 2011): Interindividual learner's characteristics and the learning task can moderate the beneficial effects of desirable difficulties on learning success. A small but growing body of research concerns this interplay; thus, one goal of the present study is to examine the role of interindividual differences in performance expectancies (PEs) prior to learning as a moderator for learning outcomes when studying with two different activities: either with problem-solving requiring greater generation activity to solve math problems, or with (a special type of) worked-examples requiring less generative activity since the solution and solution steps were explained. The explicit instructions on the solution steps decrease learners' engagement in self-explanation and therefore lower learners' generative activity, while still providing expert mental models. Our worked-examples function as comparison group to problem-solving. As such, our special worked-examples condition resembles more closely the common (re-) reading control group in research on generation (e.g., Bertsch et al., 2007) and testing effects (e.g., Kornell et al., 2012). Learning by (re-)reading can lead to overconfidence as unjustifiably high meta-cognitive judgments of one's learning compared to actual learning outcomes (e.g., Karpicke and Blunt, 2011). In this sense, studying our worked-examples may convey the (mistaken) assumption that read information is already learned, even though learners may not be able to recall the information. Such an *illusion of competence* can be the consequence of undiagnostic cues whenever information is present during studying and absent but solicited at a performance test (Koriat, 1997; Koriat and Bjork, 2006).

Desirable difficulties can decrease learners' illusion of competence (e.g., Karpicke et al., 2009; Diemand-Yauman et al., 2011) by decreasing the mismatch of cognitive processing during study and during testing (McDaniel and Butler, 2011). Test and retrieval experience in particular reduce competence illusions (Koriat and Bjork, 2006). Thus, experiencing difficulties during problem-solving as a test event requiring greater generative activity may challenge learners' competence illusion and in turn increase metacognitive accuracy (especially beyond the accuracy of studying worked-examples, which learners did not have to solve or engage in much self-explanations). In particular, literature on self-regulation has emphasized the value of accurate metacognitions for the regulation of future learning behavior (e.g., Zimmerman, 2008). Thus, another goal of the present study concerns the effects of problem-solving as the incantation of generation on changes and accuracy in PEs prompted after learning and after testing as metacognitive assessments.

Our present paper follows two related lines of argumentation. First, we introduce the generation effect as a desirable difficulty and introduce how interindividual differences can play a moderating role for learning success. These considerations serve to build the case for PEs *prior* to learning as moderators for problem-solving requiring greater generation activity than our worked-examples. We then outline how PEs *after* learning and testing may function as metacognitive assessments. These later PEs are likely differentially affected by problem-solving in contrast to worked-examples regarding competence illusions, which would pose consequences for metacognitive accuracy. Thus, PEs should be more accurate after working on greater generative problem-solving tasks than after less generative tasks of simply studying already worked-examples (with explicit explanations on solution steps).

The Generation Effect as Desirable Difficulty

The benefits of multiple desirable difficulties [e.g., generation effect, Bertsch et al. (2007); testing effect, Roediger and Karpicke (2006); distributed learning, Cepeda et al. (2006)] for memory, comprehension, and transfer are well documented (e.g., Bertsch et al., 2007; Rowland, 2014; Adesope et al., 2017). One form of desirable difficulties is the generation effect, which concerns the finding that actively generated information (e.g., solving problems, finding solutions to problems, generating answers, or producing of information) is remembered better than if the same information is more passively consumed (e.g., reading already solved problems or already worked-examples; e.g., Bertsch et al., 2007). All generative activities have in common that they require learners to engage in more effortful and deeper processing. In line with this, generated information requires learners to go beyond the information, for instance by relational processing of the provided information or by constructing links to previous knowledge (see Wittrock, 1989; Fiorella and Mayer, 2016). In line with this, actively generating information is more difficult than its mere reception (e.g., McDaniel et al., 1988; Ebbinghaus, 1913; DeWinstanley and Bjork, 2004; Bertsch et al., 2007), as is the generation of predictions and inferences rather than repetitions of solutions (e.g., Crouch et al., 2004).

Despite – or actually because of – being more difficult, self-generation can be more effective (e.g., Bertsch et al., 2007). Beneficial generation effects in learning were shown with naturalistic and/or curricular materials regarding complex topics (e.g., astronomy, engineering, physics) conducted in schools and universities (e.g., Renkl et al., 2002; Crouch et al., 2004; Richland et al., 2005; Moreno et al., 2009). Thus, positive effects of generation tasks arise in complex and realistic situations (and not only in laboratory settings using artificial or simple tasks). Furthermore, the generation effect is often thought to be related to the testing effect but considered to be broader in retrieval mode (e.g., Karpicke and Zaromb, 2010) requiring more elaborative in-depth processing (e.g., Bertsch et al., 2007; Rowland, 2014).

Moreover, the advantage of generation/testing increases for longer time periods between the generation task and the criterion test of the learned information (e.g., Bertsch et al., 2007), even though generation, for example of problem-solutions, may be undesirable in the short-term at the beginning of knowledge acquisition when worked-examples are more desirable (Kalyuga et al., 2003). Even worked-examples can outperform testing activities long-term when previous knowledge is low and the materials are high in element-interactivity (van Gog and Kester, 2012; van Gog et al., 2015). However, our special worked-examples, serving as control group, violated an important guideline (Renkl, 2014): Reducing self-explanation diminishes the effectivity of worked-examples (e.g., Berthold and Renkl, 2009; Hefter et al., 2014). The goal was to increase the difference in generative activity across both learning conditions: Problem-solving required greater generation, whereas worked-examples prompted little generative activities due to providing expert problem-solving schemes with high instructional guidance. Thus, we did not expect a worked-example effect (e.g., Schworm and Renkl, 2006; see also Wittwer and Renkl, 2010). It was necessary to avoid comparing two learning conditions that both entailed highly generative elements to examine our proposed moderation of PE and long-term effectivity for generative activities. However, worked-examples reduce cognitive load and are advantageous during initial acquisition. Problem-solving is more effective later on – after learners' expertise has increased (e.g., Renkl and Atkinson, 2003) – as well as for learners with greater previous knowledge (e.g., Kalyuga et al., 2001). This phenomenon is known as the *expertise-reversal effect* (e.g., Kalyuga et al., 2003; Kalyuga and Renkl, 2010; Spanjers et al., 2011).

Because of the difficulty of the generation task, learners can make errors while generating or fail to generate/solve problems at all (especially if they are forced to engage in such a challenging learning task; cf., Metcalfe and Kornell, 2007). The efficiency of generation, however, depends on the success of generation; more accurately, generated items lead to more learning success (e.g., Richland et al., 2005; Rowland, 2014). Thus, giving feedback and/or correcting errors moderate the benefits gained from generation tasks (e.g., Slamecka and Fevreiski, 1983; Pashler et al., 2005; Kang et al., 2007; Metcalfe and Kornell, 2007; Potts and Shanks, 2014; Metcalfe, 2017). Taking this into account, different learner characteristics potentially moderate the positive effects of generation tasks. This notion is echoed in other research (e.g., expertise reversal effect; Kalyuga et al., 2003; Kalyuga and Renkl, 2010; Spanjers et al., 2011). For instance, the expertise reversal effect states that some learning processes that prove beneficial for weaker learners or learners with lower previous knowledge (due to reduced working memory load) have no effect, or even detrimental effects, for stronger learners or learners with higher previous knowledge. Thus, it seems important to check for learner requirements or moderators that enhance the benefits of difficult learning conditions.

A hypothesis for when difficulties are desirable explicitly conceptualizes the moderating role of learners for difficulties to be desirable, specifically, the fit between learners' characteristics and the generation task; the fit of the learning content and the type of generation task; and the fit of the generation task and the performance test are interrelated (e.g., McDaniel et al., 2002; McDaniel and Butler, 2011). Thus, they emphasize learner characteristics and prerequisites as moderators for the beneficial effects of desirable difficulties on learning success.

On the one hand, the authors (e.g., McDaniel et al., 2002; McDaniel and Butler, 2011) imply that desirable difficulties may be especially beneficial for learners with lower (cognitive) abilities. That is, difficulties could lead to cognitive processes and applications of effective strategies that learners would not have spontaneously used themselves. This in turn enhances learning, so desirable difficulties instigate compensatory processes. For instance, different studies implementing varying forms of desirable difficulties supported this assumption for the following abilities: lower general intelligence, lower structure building readers, and lower cognitive motivation (lower need for cognition; McDaniel et al., 2002; Brewer and Unsworth, 2012; Schindler et al., 2019).

On the other hand, researchers also implied that desirable difficulties can only increase learning if learners are able to fulfill the prerequisites of the difficult tasks. Hence, the effectivity of the desirable difficulties is tied to complementary preconditions between learners and tasks. For instance, studies showed higher previous knowledge and higher reading skills to be prerequisites for beneficial desirable difficulties (McNamara et al., 1996; McDaniel et al., 2002). McDaniel et al. (2002) supposed that less able readers had to use most of their processing capacities to correctly generate the items, so that they had no cognitive resources left to further process and encode the information.

These assumptions indicate that learner characteristics can moderate the beneficial effects of desirable difficulties in the above-mentioned two ways. However, the assessment of learner characteristics has so far not been exhaustive, meaning that further characteristics, for instance (cognitive-motivational) expectancies, are worthy to be explored.

Performance Expectancies Prior to Learning as Moderator for the Generation Effect

One such learner characteristic worth examining may be performance expectancies (further PE/PEs). Expectancies are theorized to influence learners' behavioral orientations as well as the intensity or persistence of learners' behaviors and consequently their performance (e.g., Eccles, 1983; Eccles and Wigfield, 2002). PEs describe individuals' subjective beliefs or ratings of how well one will perform in academic or achievement related tasks (e.g., Eccles, 1983; Eccles and Wigfield, 2002; Marshall and Brown, 2004) and could be related to or influenced by previous knowledge (for instance, higher previous knowledge could enhance the expectation to solve the same tasks). PEs are metacognitive predictions about future performances with motivational consequences: Such expectancies have been shown to be positively related to actual performance because they can shape the time and effort learners invest in tasks (e.g., Marshall and Brown, 2004; Schindler et al., 2016). PEs depend on an individual's self-concept and the perceived difficulty of the learning task (e.g., Marshall and Brown, 2004; Dickhäuser

and Reinhard, 2006). PEs only enhanced actual performance for difficult tasks but had no influence on performance on easy tasks [probably because these can be solved without further effort; e.g., Marshall and Brown (2004), Reinhard and Dickhäuser (2009)]. This should be especially relevant for desirable difficulties, which are inherently more difficult learning tasks.

Accordingly, generation tasks (and the required more intensive and deeper information processing) should be more effective for learners with higher PEs: Learners with higher PEs should better match the difficult generation tasks because they are more motivated to exert (cognitive) effort, time, and persistence. In contrast, low PEs can potentially reduce learners' motivation and persistence while working on generation tasks because learners believe that they will not be able to solve the tasks. Further, higher PEs can be seen as a more relevant learner characteristic for (difficult) tasks in which participants must actually solve problems, in contrast to (easier) tasks in which they have to read worked-examples.

Performance Expectancies After Learning and Testing as Metacognitive Assessments

The previous considerations focused on PEs – formed prior to working on a learning task – as a learner characteristic, which may function as a moderator for learning success. PE in this sense is identified as another potential moderator similar to other moderators discussed above, like previous knowledge. The difference of PE in comparison to these aforementioned moderators lies in the metacognitive nature of PE, whereas previous knowledge is cognitive in nature. Thus, a metacognitive performance judgment prior to learning may moderate learning success. This can be seen as one part of the story. The second part concerns how metacognitive judgments can act as a moderator for regulatory processes during and after learning, and therefore act as a mediator for learning success (e.g., Serra and Metcalfe, 2009). In this sense, PE – formed during or after learning and testing – may potentially be tied to metacognitive accuracy and metacognitive accuracy (in tandem with regulation accuracy) was shown to function as a mediator for learning success (e.g., Thiede et al., 2003). Therefore, we will briefly consider how solving problems opposed to studying problem solutions may influence metacognitive assessments and accuracy.

Metacognition – which refers to the knowledge of one's own cognitive processes – can direct regulatory processes such as restudy choices (Dunlosky and Metcalfe, 2009). For example, problem-solving can improve the accuracy of judgments of learning (JOL) by decreasing performance overestimations (Baars et al., 2014, 2016). Accurately estimating and monitoring one's performance are important educational outcomes because accurate metacognition effectively guides studying (Dunlosky and Lipko, 2007). Since metacognitive assessments guide learning, for example, by invested time (Son and Metcalfe, 2000), mental effort (Mihalca et al., 2017), or restudy decisions (Thiede et al., 2003; Dunlosky and Rawson, 2012), so do PEs influence time and effort allocations (Schindler et al., 2016). Since PEs describe individuals' subjective performance beliefs (Marshall and Brown, 2004), they are (task-specific) metacognitive competence ratings and as such are a form of metacognitive judgment. PEs prompted after learning – similar to JOL prompted after learning – should be less influenced by the self-concept and instead should be more rooted in the experience of the actual learning task. Therefore, previously found effects of problem-solving versus worked-example studying on metacognitions and accuracy are likely to apply to PEs as well.

Effects of Problem-Solving on Performance Expectancies as Metacognitive Assessments

In contrast to problem-solving, worked-examples can be seen as procedural solution scaffolds and are thereby mentally less taxing (in terms of working-memory load) and designed to ease schema construction (Sweller, 2006). However, such reduced difficulty (relative to problem-solving) can have metacognitive drawbacks in terms of conveying an illusion of competence after studying worked-examples (Baars et al., 2014, 2016), for example, when the content is currently accessible but will not be (completely) available later (e.g., Koriat and Bjork, 2005; Karpicke et al., 2009). Competence illusions during and after learning can negatively impact learning success: Overconfidence can lead to faulty regulation, such as early study termination or inaccurate selection of materials for restudy (Thiede et al., 2003; Dunlosky and Rawson, 2012). Overconfidence may also lead learners to underestimate the effort necessary to internalize correct and complete problem-solving schemas from worked-examples (Kant et al., 2017). Thus, experiencing difficulties while learning with problem-solving may challenges learner's competence illusion, which may stimulate learners to engage in deeper and (cognitive) more effortful information processing (e.g., McNamara et al., 1996; Diemand-Yauman et al., 2011); and increase metacognitive accuracy in terms of predicted performance and actual performance (Baars et al., 2014); as well as, increase regulation accuracy in terms of selecting the right materials for restudy (Baars et al., 2016).

Multiple reasons are discussed as to why problem-solving can improve metacognitive accuracy. Baars et al. (2016) suggests that problem-solving as a generation activity allows learners to recall and test the quality of their acquired schema. They further capitalize on the idea of postdiction judgments (Griffin et al., 2009), which refers to the idea of utilizing test performance of a previously completed task as a cue on which to base judgments. Others suggest that encoding and retrieval fluency can influence metacognitive judgments (Agarwal et al., 2008; Pieger et al., 2017). All have the same implication that problem-solving entails more accurate cues on which to base metacognitive judgments, reducing overconfidence and increasing metacognitive accuracy (Kant et al., 2017).

The presented logic and previous findings of problem-solving versus studying worked-example on metacognitions and accuracy should also prove applicable to PEs: Learners may use the experienced difficulty of solving problems as opposed to reading worked-examples as a cue to lower their PEs, because the

difficulty of solving problems may challenge learners' competence illusion. In contrast, reading less difficult worked-examples may not challenge learners' competence misconceptions. If so, learners in the problem-solving condition should decrease their PEs after the learning task and indicate more accurate PEs with respect to the later test outcome. Learners in the worked-example condition should not adjust their PEs. Hence, metacognitive accuracy should be improved in the problem-solving condition in contrast to worked-examples.

The Present Study

The present work focuses on the generation effect and examines the potentially moderating role of learners' initial PEs. Generation tasks are demanding tasks that require the recruitment of more cognitive capacities and deeper/more elaborate processing to solve the tasks and overcome the challenge. Thus, learners must exert more thinking, more time, and more effort to solve such tasks to reap their benefits. Hence, participants should be motivated and persistent, but this is not automatically the case for every learner. Regarding learner characteristics, PEs can lead to higher performance in achievement tasks through more allocation of resources like time, persistence, and effort. Thus, higher PEs can be seen as a fit between the generation tasks and learners' abilities to cope with them. As mentioned above, a better fit between (cognitive) prerequisites of the task and (motivational-cognitive) characteristics of the learner is important for the effectiveness of such difficulties. Learners with higher PE are potentially more prone to exert and persist in more effortful processing.

Due to the above theoretical and empirical arguments, we propose the following hypotheses: (H1) We assume a two-way interaction between the condition (problem-solving vs. worked-example) and time (immediate vs. delayed). Performances in the worked-example condition should be higher in the immediate test, while the performances benefits of problem-solving should be apparent at the delayed test (time × condition). We also suppose a two-way interaction between the condition and PEs. (H2) Higher PEs should be more advantageous when solving problems compared to reading worked-examples (PE × condition). Since generation effects are desirable difficulties that often have greater delayed benefits rather than immediate benefits, we can assume a three-way interaction of condition × PE × time. (H3) The advantage of problem-solving for higher PEs should be more pronounced later in the delayed performance test rather than in the immediate test. Therefore, we predict a three-way interaction of PEs, condition, and time (PE × condition × time). We tested these hypotheses based on students' immediate and delayed performance (after 3 months) using coherent curricular materials on linear functions in a sample of eighth graders (advanced school track) and measuring PEs prior to engaging in the learning task.

The present work also investigates the effects of problem-solving and worked-examples on PEs after learning as a competence-related form of metacognitive judgment. Since problem-solving can affect metacognitive assessments and accuracy by decreasing competence illusions, the difficulty of solving problems may challenge a learner's initial performance overestimates. In contrast, a mere reading of problems and their solutions should align with a higher (misplaced) sense of competence (cognitive illusion), which should result in higher PEs for problem-information than for read-only.

We will thus test if the formation of more accurate PEs depends on active problem-solving required by the learning task: Initial PE prior to learning (and hence prior the experimental manipulation) should not differ, whereas during learning (and hence depending on the experimental learning condition), PEs in the solving condition should be lower compared to the worked-example condition. This difference should be eliminated once the problems of the performance test are completed by all (that is, also by worked-example learners), and pupils must indicate retrospectively how well they thought they did in the test (because all learners experienced the difficulty of problem-solving, in this case of the test problems).

Regarding later PEs prior to the second performance test 3 months later, it is possible that pupils base their PEs on their judgments of their performance after the first test. In this scenario, PEs prior to the second test may equal the post-test PEs. Another scenario may be that pupils remember the learning experience and base it on the experienced difficulty while learning, thus PEs prior to the second test may be lower in the problem-solving condition. In either case, we predicted an interaction effect of condition and time on metacognitive judgments of performance (H4). Moreover, calibration accuracy (a smaller difference between expected performance and actual performance) should be more precise for problem-solvers in contrast to worked-examples: If learners in the problem-solving condition decrease their PEs after the learning task, their PEs should be more accurate with respect to the later test outcome. Learners' unadjusted PEs in the worked-example condition maintain a competence misconception and therefore should be less accurate (H5). We tested these hypotheses by additionally measuring PEs after learning, after the immediate performance test, and prior to the delayed performance test.

MATERIALS AND METHODS

Participants and Design

Participants were children in the eighth grade of the secondary school track recruited from a school located in a medium-sized town in Germany. Written, full, informed consent was obtained from the principals, teachers, parents, and children,[1] which resulted in an initial sample of $N = 71$. Not all participants were present at the first in-class session in school, nor at the second in-class session 3 months later, resulting in $n_{\text{Session 1}} = 68$ (41 females) and $n_{\text{Session 2}} = 64$ (39 females). This led to $n = 61$ pupils being present at both in-class sessions (32 in the worked-example condition and 29 in the problem-solving condition; mean

[1]This study was conducted in full accordance with the Ethical Guidelines of the German Association of Psychologists (DGPs) and the American Psychological Association (APA). This study was fully approved by the Ethics Committee affiliated with the Hessian Ministry for the Science and the Arts (Yasar Karakas, Hessisches Kultusministerium, Referat I.3.2, Luisenplatz 10, 65185 Wiesbaden, Germany; Phone: +49 611 368 – 2734; E-mail: Yasar.Karakas@kultus.hessen.de).

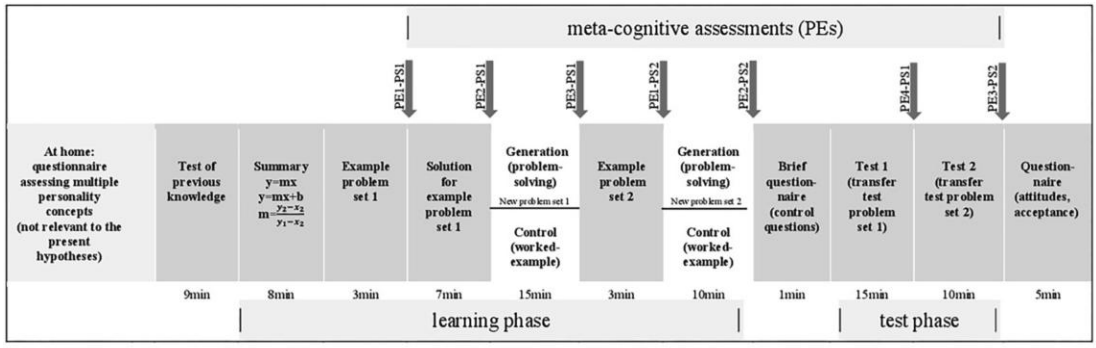


FIGURE 1 | Timeline and schematic design in-class Session Time 1. Gray-colored arrays denote the same procedures and materials for all participants; white arrays show the differing procedures and materials according to the experimental manipulation. PE, performance expectancy; PT, problem set; EP, estimated performance; thus PE1-PS1, performance expectancy measurement 1 for problem set 1; PE2-PS1, performance expectancy measurement 2 for problem set 1; PE3-PS1, performance expectancy measurement 3 for problem set 1; PE1-PS2, performance expectancy measurement 1 for problem set 2; PE2-PS2, performance expectancy measurement 2 for problem set 2; EP1-PS1, estimated performance for problem set 1; EP1-PS2, estimated performance for problem set 2.

age = 13.64 years, *SD* = 0.58). The participants were randomly assigned to either the experimental condition (problem-solving) or control condition (worked-examples). The randomization was successful as the condition was not related to gender distributions ($\Phi_{T1} = 0.10$, $p = 0.46$) or to competence distributions indicated by previous math grade (Spearman's $\rho = -0.14$, $p = 0.30$; $M_{\text{worked-example}} = 3.38$ (∼C), $SD = 1.04$; $M_{\text{problem-solving}} = 3.03$ (∼C), $SD = 1.12$; $F(1,59) = 1.52$, $p = 0.22$, $\eta_p^2 = 0.03$), or to previous knowledge (Point-biserial $r = 0.05$, $p = 0.70$; $M_{\text{worked-example}} = 10.95$, $SD = 3.93$; $M_{\text{problem-solving}} = 11.32$, $SD = 3.48$; $F(1,59) = 0.15$, $p = 0.70$, $\eta_p^2 = 0.00$).[2] Thus, pupils in both conditions had similar prerequisites. The materials were pre-tested and adapted in a (different) sample of $n = 30$ eighth graders prior to administration of the materials in their final form in the current sample. The study was a 2(condition: solving vs. worked-examples) × 2(post-test time point: immediate vs. delayed) design with condition as between-subjects factor and post-test time-point as within-subjects factor. As a token of appreciation at the end of the study, the children received sweets and a small gift (puzzles) for their time and effort.

Procedures

Prior to the study, the teachers had briefly introduced the topic of linear functions to the children. The children were novices and therefore they had very low previous knowledge. The teachers were instructed to omit any exercises that would be related to computing slopes and functions in their introductory teachings. Furthermore, the teachers handed short questionnaires to the children. They contained the measurements of multiple personality variables (not relevant to the proposed hypotheses in this paper but covered in another manuscript on the relationship of personality and long-term performance in a surprise test) and were collected by the researchers prior to the in-class session. All obtained data from the participants were pseudonymized based on number codes to allow subsequent matching of the data in both in-class sessions. Data collection in all school classes was conducted by the second author, supported by research assistants. All materials were paper–pencil contained in folders. All participants were allowed to use a calculator.

First In-Class Session

In class, participants were randomly re-seated. Multi-colored maps veiled the study's condition to the children. The color-code also served to avoid having children with the same experimental condition clustered together. After a brief welcome to the pupils, all instructions were scripted, and all activities were timed. **Figure 1** shows the procedures schematically. After a short test on participants' previous knowledge (see **Appendix Figure A1** in **Supplementary Material**), participants received two explanatory content pages (see **Appendix Figure A2** in **Supplementary Material**). All participants were instructed to read the contents carefully, to try to comprehend them, and to keep in mind the important information highlighted in bold, bright red. They were told repeatedly that they would need the highlighted information in these explanatory materials for the upcoming test.

Once participants had studied the explanatory materials in their folders, they received a brief example of test problem set 1 (see **Appendix Figure A3** in **Supplementary Material**). It had the same surface structure as in the upcoming test. Due to that, participants were asked to indicate their PEs for this problem set (PE1-PS1). Subsequently, all participants received the correct solution steps for this particularly presented example problem, followed by a second assessment of their PEs (PE2-PS1).

The subsequent pages contained further problems of set 1, yet these problems differed for the participants depending on the experimental condition they were in (see **Figure 2**). In the worked-example condition, participants received these problems with all correct solution steps worked out, accompanied by short explanations of the steps; in the solving condition, participants had to solve all problems by themselves, however, they could refer to the previous materials (open-book). The instruction for

[2]The pupils were novices to linear functions. Most points in the previous knowledge test were achieved based on recognizing the graphs of linear functions (**Appendix Figure A1** in **Supplementary Material**, task 2); not by the tasks 4–7.

Solving Condition

a) (0|0); (2|2) fits the [ ] line and has the functional equation [ ] with the slope m = [ ]; the point (22|22)

[ ] is [ ] is not on this line, because [ ].

Worked-Example Condition

a) (0|0); (2|2) fits the [blue] line and has the functional equation [y = x] with the slope m = [1]; the point (22|22)

[X] is [ ] is not on this line, because [$\frac{22}{22} = 1$ or $22 = 1 \cdot 22$].

Correct solution:

Step 1: Insert the given points (2|2) and (0|0) into the formula to compute the slope.

$m = \frac{y}{x} = \frac{2}{2}$ | Recap, if you enter (2|2) and (0|0) in the general formula to compute the slope $m = \frac{y_2 - y_1}{x_2 - x_1}$, then you will obtain $= \frac{2-0}{2-0} = \frac{2}{2} = \frac{y}{x}$

$m = 1$

Step 2: This computed slope of m = 1 is inserted into the equation defining the line through the origin of y = m·x.

$y = 1 \cdot x$ | Insert $m = 1$ into $y = m \cdot x$

$y = x$

Step 3: Enter the new unknown point (22|22) with a known point on the line, e. g. (0|0) or (2|2) into the formula to compute the slope.

$m = \frac{22-0}{22-0} = \frac{y}{x} = \frac{22}{22} = 1$ *or* $m = \frac{22-2}{22-2} = \frac{20}{20} = 1$

Step 4: Compare, if your computed slope based on an unknown and known point in step 3 matches the previously computed slope in step 1.

$m = 1$ stemming from the functional equation of $y = 1 \cdot x$ in step 1

$m = \frac{22}{22} = 1$ stemming from the computed slope of an unknown and known point in step 3

Both slopes of $m = \frac{22}{22} = 1$ and of $m = 1$ are equal, that is $1 = 1$. Therefore, the unknown point has to fit the blue line.

alternatively:

Enter the unknown point of (22|22) into your derived functional equation from step 2 and check, whether the result is a true mathematical statement.

$y = m \cdot x$ | Point (22|22) has $x = 22$ and $y = 22$

$22 = 1 \cdot 22$

$22 = 22$ | This is true. Thus, the unknown point (22|22) has to fit the blue line.

FIGURE 2 | Learning phase: manipulation problem set 1.

participants in the worked-example condition read, "Please read the correct solution steps thoroughly, try to comprehend them, and learn them." The instruction for the solving-condition read, "Please try to solve all problems." Participants in the solving-condition were provided with the correct solutions for 2 min at the end of this task, then all participants were asked a third time to indicate their PEs for this problem set (PE3-PS1).

The next page contained an example of a new problem set (problem set 2), which had a similar surface structure as in the upcoming test (see **Appendix Figure A4** in **Supplementary Material**). Due to that, participants were asked to indicate their PEs (PE1-PS2). Again, the following pages differed for participants depending on their experimental condition (see **Figure 3**). In the worked-example condition, the correct solution steps were displayed with some explanations. In the solving conditions, the problem set had to be solved. Again, participants in the solving condition could use the previous materials as reference to help them solve the task, and they were provided with the correct solution for 2 min. For all participants, the next page contained the second PE measurement of problem set 2 (PE2-PS2). A short survey with control questions (like perceived task difficulty or invested effort) concluded the learning phase prior to the test phase (see **Appendix B** in **Supplementary Material**).

Participants started the test phase with seven new problems of set 1 and 15 min time to solve them (see **Appendix Figure A5** in **Supplementary Material**). Thereafter, participants had 30 s to estimate how well they had just performed (PE4-PS1). Participants then continued with new problems of set 2 and 10 min of time and afterward were asked once again to estimate how well they had just performed (PE3-PS2). Once the test phase was finished, they answered questions regarding their overall learning and test experience and about their attitudes toward the learning method.

Second In-Class Session

Both in-class sessions were 3 months apart (see **Figure 4** for the schematic design of In-class Session 2). As in the previous session, participants were randomly re-seated. Once participants had opened their folders, they read that they would receive the exact same set of test problems as in Session 1 (pupils did not expect the second test). Yet, prior to the second test, they were again asked to indicate their PEs for the test problem set 1 (PE5-PS1) and for the test problem set 2 (PE3-PS2). Thereafter, pupils had 15 min time for the test problems of set 1 and 10 min time for the test problems of set 2. (Then, pupils had 20 min to solve the new surprise test problems, which were irrelevant to the hypothesis tested.) Session 2 concluded with a brief questionnaire with control questions (e.g., whether they took the test seriously and how much effort they invested in solving the problems). Finally, participants were thanked and dismissed.

Materials and Measurements

Given the complexity of the materials, a few words on the materials' structure and logic is warranted. The materials represented real curricular contents and were developed in cooperation with subject didactics. The contents focused on linear functions, specifically on computing slopes and deriving equations. Of the explanatory material (see **Appendix Figure A2** in **Supplementary Material**), the first page pointed out similarities between a bijective mapping rule and the equation of a positive linear function. Both forms, $y = mx$ (through the origin) and $y = mx + b$ (shifted origin) were covered. The second page contained new content for the participants: the logic behind a slope and its formula for computation, the logic behind the y-axis and the constant b, and the link to the equation of a linear function. The materials of both problem sets in the learning phase focused on positive linear functions through the origin. Both test problems (see **Appendix Figures A5**, **A6** in **Supplementary Material**) required transfer to negative linear functions. Both forms were required ($y = mx$; $y = mx + b$).

Problem Sets

Two coherent problem sets were chosen. Therefore, all following measurements and manipulations had to be phrased for both problem sets. For the analyses, like in any exam, one final score represented the test performance comprised of both problem sets.

Problem set 1

Problem set 1 required of the participants to (a) identify a line based on two given points in a coordinate system; (b) derive the functional equation; (c) compute the slope; (d) indicate whether a new point lies on the same line; and (e) proof the answer mathematically. Problem set 1 focused more on the execution of arithmetic computational procedures based on abstract contents.

Problem set 2

Problem set 2 required (a) sketching of a graph into a coordinate system; (b) finding a specific y-value in the graph; (c) explaining what a slope is; (d) computing the slope; (e) deriving the functional equation; and (f) computing a specific x-value. Problem set 2 focused more on the application of arithmetic formula to real-world contents.

Performance Expectancies

PEs were assessed as task-specific and therefore measured separately for each problem set (see **Figure 1**). After participants were shown an exemplary test problem of set 1 (see **Appendix Figure A3** in **Supplementary Material**), three items recorded their PEs. The first item read, "*How well do you think you will perform in the upcoming test with this type of problems? Please estimate which grade you will be able to achieve in a test with seven test problems of this type.*" The range is from 1 = very good [A] to 6 = fail [F]). The second item read, "*How many points of 35 total do you think you will be able to achieve in the upcoming test?*" The range is from 0 to 35. The third item read, "*How many of the seven test problems of this type do you think you will be able to solve correctly in the upcoming test in 15-minutes of time?*" The range is from 1 to 7.[3] PEs for problem set 2 were measured with two items (see **Appendix Figure A4** in **Supplementary Material**).

[3]Note that we included three questions to assess performance expectancies because we were not sure about accuracy and variance of pupils' judgments – whether eighth graders would naturally judge their performance in expected grades, or points, or number of test problems solved – and whether grade and number of test problems solved would vary enough for meaningful analyses. Since the three assessments were highly correlated and variances were highest for judgments in points, we included the performance expectancies of points in the main analyses. (Expected points also had the same metric as points achieved in both post-tests.)

Solving Condition

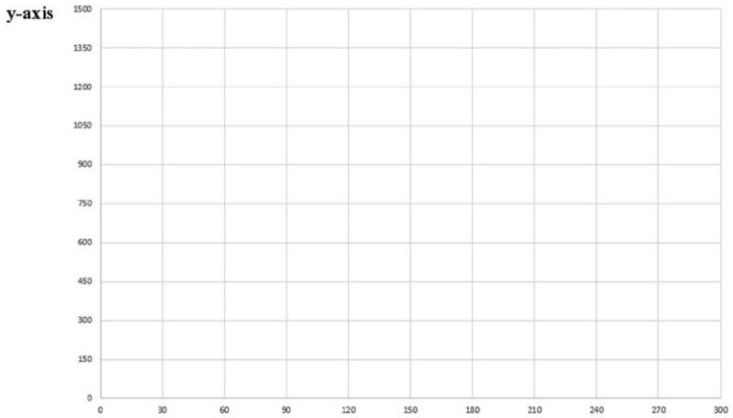
If you need help, you can refer to the pages „From mappings to functions" in your folder.

Exercise 2: Complete accordingly.

If you watch a dripping water tap for a continued period of time, then you will notice, it is dripping consistently. Susanne made the following measurements:

Time since the water tap drips in min	**0**	**60**	**90**	**240**
Amount of water in the sink in ml	**0**	**300**	**450**	**1200**

a) In the graph, **indicate** how much water has been dripping into the sink after 150 minutes. **Chart** the corresponding line in the coordinate system, **write down** this number.



Worked-Example Condition

Exercise 2: Read and memorize the solution steps.

If you watch a dripping water tap for a continued period of time, then you will notice, it is dripping consistently. Susanne made the following measurements:

Time since the water tap drips in min	**0**	**60**	**90**	**240**
Amount of water in the sink in ml	**0**	**300**	**450**	**1200**

b) In the graph, **indicate** how much water has been dripping into the sink after 150 minutes. **Chart** the corresponding line in the coordinate system, **write down** this number.

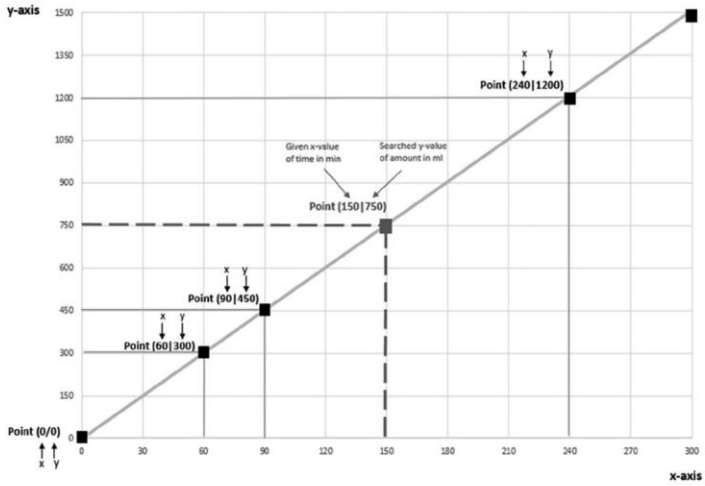


FIGURE 3 | Learning phase: manipulation problem set 2.

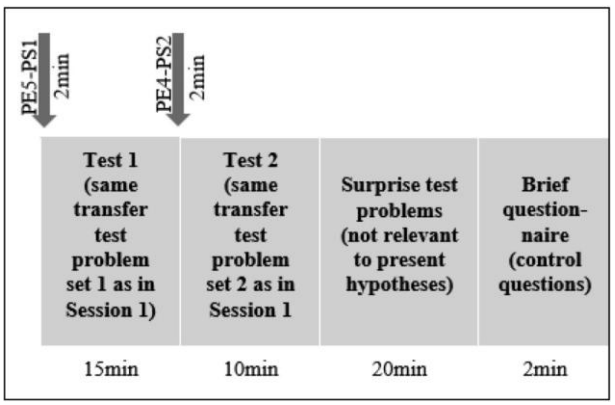


FIGURE 4 | Timeline and schematic design in-class Session Time 2. PE4-PS1, performance expectancy measurement 4 for problem set 1; PE1-PS2, performance expectancy measurement 3 for problem set 2.

In the test phase, after completing each test problem set, participants were asked to retrospectively estimate how well they *had* performed. The item for one's post-test PEs (problem set 1) read, "*How well do you think you performed with respect to the previous test problems?*" Possible answers included, "*I think that I achieved _____ (grade).*"; "*I think I solved ____ (number) of seven problems correct*"; "*I think, I obtained ____ (points) of 35 points*." The item for one's post-test PEs of set 2 mirrored the items for problem set 1 (without the third item).

Experimental Manipulation

Figures 2, **3** illustrate the difference between both experimental conditions in the learning phase. In the Solving Condition, seven different problems of set 1 had to be solved. In the Worked-Example Condition, the same seven problems were presented along with their correct solutions, and along with each step necessary to solve the problem correctly (including short explanations). Likewise, in the Solving Condition, problem set 2 had to be solved by working alone, while in the worked-example condition the solutions and step-by-step guidance were provided. The instructions differed accordingly: "*Read, try to comprehend, and learn them*," versus "*work out the solution by yourself*." Participants in the solving conditions received the correct solutions to both problems for comparison.

Test Problems

The test problem sets had the same surface structure as the problems sets in the learning phase but required transfer (the problem sets for instance included only positive slopes and point of origins in (0| 0), whereas the slopes in the test problems were also negative and the points of origins could differ). **Appendix Figures A5**, **A6** in **Supplementary Material** display all used test problems. The same test problem sets were used in Sessions 1 and 2. Two independent raters coded pupils' answers to the test problems with high interrater-reliability (Session 1 $r = 0.95$ and Session 2 $r = 0.97$). Any remaining discrepancies were discussed and resolved. A total of 42 points could be achieved (with 35 points for problem set 1 and 7 points for problem set 2); Cronbach's $\alpha = 0.88$ (immediate post-test), Cronbach's $\alpha = 0.92$ (delayed post-test).

RESULTS

Performance by Learning Conditions Across Post-tests

Exercising with worked-examples should be superior to problem-solving with respect to an immediate performance, but inferior to problem-solving in a later performance test (H1; see **Table 1** for descriptive statistics). An rANOVA with time as within-subject factor and condition as between-subject factor (0 = worked-examples, 1 = problem-solving) tested this proposition. We found a main effect of time, $F(1,59) = 9.34$, $p = 0.003$, $MD = -2.31$, $SE = 0.75$, 95% CI [−3.81, −0.80], $\eta_p^2 = 0.14$, which means that the overall performance worsened by about 2 points. We found no main effect of condition, $F(1,59) = 2.57$, $p = 0.11$, $MD = 2.43$, $SE = 1.52$, $\eta_p^2 = 0.04$, 95% CI [−0.60, 5.47], only descriptively performances in the problem-solving condition ($M = 15.75$, $SE = 1.10$, 95% CI [13.55, 17.94]) was 2.43 points higher than in

TABLE 1 | Descriptive statistics of the central variables.

Variable	Condition			
	Worked-examples		Problem-solving	
	M(SD)	95% CI	*M(SD)*	95% CI
In-class Session 1				
Initial performance expectancy in points[4]	23.20 (7.94)	[20.27; 25.98]	18.87 (10.09)	[15.30; 22.35]
Test performance in points	14.13 (5.30)	[12.39; 16.08]	17.24 (6.85)	[15.07; 19.69]
In-class Session 2				
Test performance in points	12.50 (6.58)	[10.49; 14.90]	14.25 (7.62)	[11.81; 16.93]

95% CI is based on bootstrapping with 1000 samples. All initial performance expectancies and both post-test points ranged from 0 to 42 points. [4]When Initial performance expectancy is computed as mean of PS1-PE1 and PS1-PE2, summed with PS2-PE1 (that is with the initial performance expectancy after receiving an example solution averaged), the values are similar, M = 23.54, SD = 8.01, 95% CI [20.70, 26.19]. Performance expectancies prior and after seeing the example solution (PS1-PE1 and PS1-PE2) are not statistically different in both groups (Worked example: t(31) = −0.97, p = 0.34, MD = −0.67, SD = 3.91, 95% CI [−2.08, 0.74], $d_{Cohen} = 0.09$; Problem solving: t(28) = −1.45, p = 0.16, MD = −1.17, SD = 4.37, 95% CI [−2.83, 0.49], $d_{Cohen} = 0.13$).

the worked-examples condition (M = 13.31, SE = 1.05, 95% CI [11.22, 15.41]). We obtained no interaction of time × condition, F[1, 59] = 0.83, p = 0.37, B = −1.37, SE = 1.51, 95% CI [−4.39, 1.6], η_p^2 = 0.01. Thus, there is no support for the proposed 2-way interaction of condition and time (H1).[4]

Performance by Learning Conditions Across Post-tests Moderated by Performance Expectancies

The following analyses scrutinize whether the effectivity of both learning conditions differed as a function of post-test time point and (standardized) initial PEs (sum of PS1-PE1 and PS2-PE1). We examined whether learning with problem-solving was better for pupils with higher PEs (H2), especially in the long run (H3). All tests are reported two-tailed; the follow-up analyses as mean comparisons are conducted within the subsequent model and, if necessary, considered for higher (+1SD) and lower (−1SD) levels of standardized initial PEs and complemented by regions of significance (Johnson-Neyman technique; determined with PROCESS, Hayes, 2018).

We conducted repeated measures analyses of variance with time as within-subjects variable, condition as between-subjects variable (0 = worked-examples, 1 = problem-solving), and the standardized initial performance expectancy as a continuous moderator (cf. Judd et al., 2001) to specify the two-way and three-way interactions. We were predicting a two-way interaction of time × condition (H1), a two-way interaction of condition × initial performance expectancy (H2), as well as a three-way interaction of time × condition × initial performance expectancy (H3).

The results show a main effect of time, F(1,57) = 13.26, p < 0.001, MD = −2.70, SE = 0.74, 95% CI [−4.19, −1.22], η_p^2 = 0.19, a main effect of initial performance expectancy, F(1,57) = 19.83, p < 0.001, η_p^2 = 0.26, and a main effect of condition, F(1,57) = 8.17, MD = 3.89, SE = 1.36, p = 0.006, 95% CI [1.17, 6.52], η_p^2 = 0.13. Again, we did not obtain the expected interaction of time and condition (H1), F(1,57) = 0.24, p = 0.62, MD = −0.73, SE = 1.48, 95% CI [−3.70, 2.24], η_p^2 = 0.00. We found no convincing evidence for an interaction of initial performance expectancy and time, F(1,57) = 3.62, p = 0.06, η_p^2 = 0.06, and we did not find the predicted interaction of initial performance expectancy and condition (H2), F(1,57) = 0.08, p = 0.93, η_p^2 = 0.00; nevertheless, the postulated three-way interaction of time, initial performance expectancy, and condition was significant (H3), F(1,57) = 5.30, p = 0.025, B = −3.50, SE = 1.52, 95% CI [6.54, 0.46], η_p^2 = 0.09.[5]

To understand these findings, we first attend to the adjusted main effects (for pupils with average initial PEs), which can be interpreted as performance decreases across time by about 2.5 points. The higher the initial PEs, the better pupils performed, and the overall performance in the problem-solving condition was about 4 points higher than in the worked-example condition. Note that the main effects of time and condition are the adjusted effects under consideration of initial PEs and thus represent the effects for an average level of initial PEs. The middle of **Figures 5–7** illustrates these time and condition effects. More specifically (and given an average level of initial PE), in the immediate post-test, pupils in the problem-solving condition achieved 4.26 point more than those in the worked-example condition, MD = 4.26, SE = 1.47, p = 0.005, 95% CI [1.31, 7.19], Cohen's d = 0.76, which amounted to a 3.52 point

[4]The results are the same when including (standardized) previous knowledge or math grade in the model.

[5]The results are the same when including (standardized) previous knowledge or math grade in the model. Thus, both previous knowledge and previous math grade, that correlate with performance expectancies and post-test performance, can be ruled out as alternative explanations. The results are also similar when computing the model with initial performance expectancies as mean of PS1-PE1 and PS1-PE2, summed with PS2-PE1.

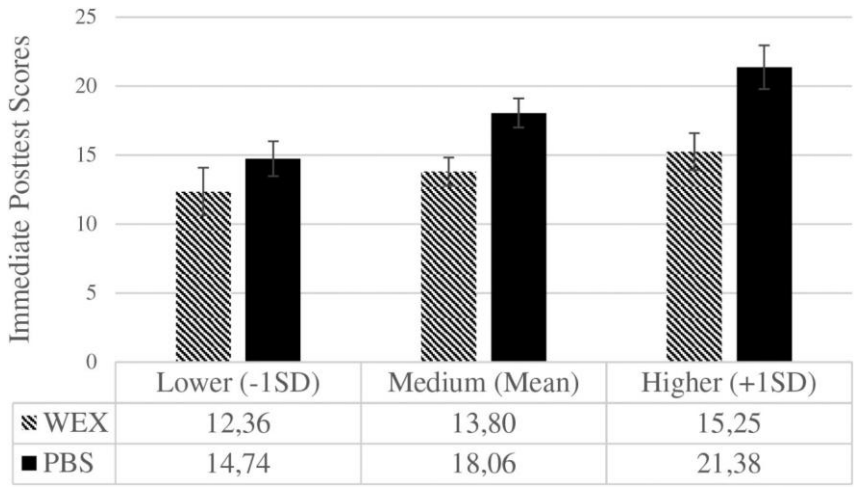


FIGURE 5 | Immediate post-test scores for both learning conditions at different levels of initial performance expectancies. WEX, worked-examples (0), n = 32; PBS, problem-solving (1), n = 28. Error bars represent the standard error of the mean [WEX: 1.73 (–1SD), 1.02 (Mean), 1.33 (+1SD); PBS: 1.26 (–1SD), 1.06 (Mean), 1.58 (+1SD)]. Performance expectancies (standardized) are depicted for lower, medium, and higher levels. Post-test scores could range from 0 to 42.

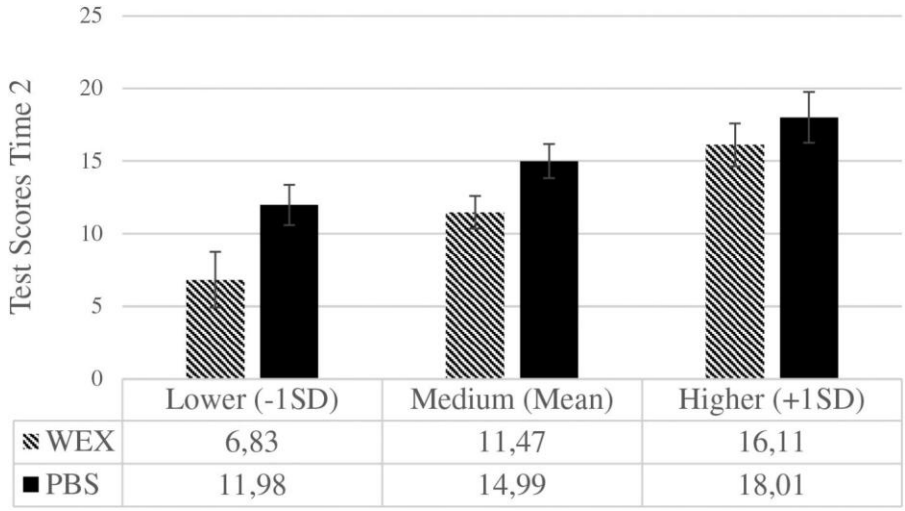


FIGURE 6 | Delayed post-test scores for both learning conditions at different levels of initial performance expectancies. WEX, worked-examples (0), $n = 32$; PBS, problem-solving (1), $n = 29$. Error bars represent the standard error of the mean [WEX: 1.91 (–1SD), 1.13 (Mean), 1.48 (+1SD); PBS: 1.40 (–1SD), 1.17 (Mean), 1.75 (+1SD)]. Performance expectancies (standardized) are depicted for lower, medium, and higher levels. Post-test scores could range from 0 to 42.

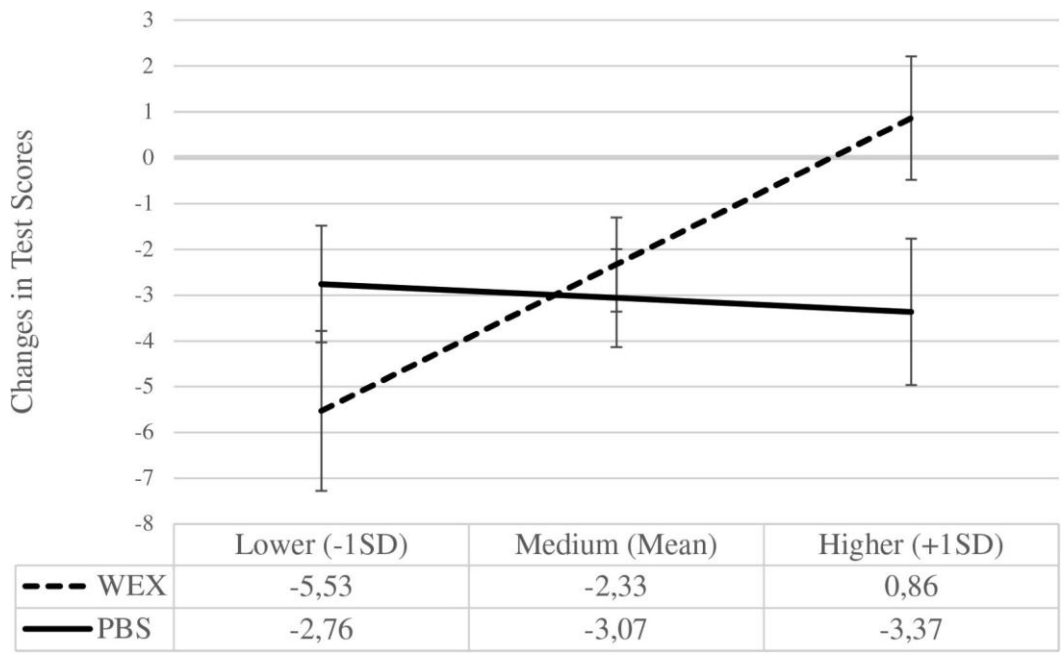


FIGURE 7 | Performance changes across both post-test by learning condition and initial performance expectancies. Change scores on the *y*-axis were computed by subtracting the delayed post-test scores from the immediate post-test scores: Zero means no change, negative values mean performance loss, and positive values mean performance gains. The *x*-axis anchors these changes for both learning conditions (WEX, worked-example (0), $n = 32$; PBS, problem-solving (1), $n = 29$) for lower, medium, and higher levels of performance expectancies. Error bars represent the standard error of the mean [WEX: 1.74 (–1SD), 1.03 (Mean), 1.35 (+1SD); PBS: 1.27 (–1SD), 1.07 (Mean), 1.60 (+1SD)].

advantage in the delayed post-test, $MD = 3.52$, $SE = 1.63$, $p = 0.034$, 95% CI [0.27, 6.78], Cohen's $d = 0.56$. The lack of support for the time × condition interaction is due to statistically similar performance decline over time in both learning conditions, $MD = -0.73$, $SE = 1.48$, 95% CI [−3.70, 2.23], Cohen's $d = -0.13$. In the worked-example condition, post-test performance decreased about 2.5 points over time, $MD = -2.33$, $SE = 1.03$, 95% CI [−4.39, −0.28], Cohen's $d = -0.39$, but about 3 points in the problem-solving condition $MD = -3.07$, $SE = 1.07$, 95% CI [−5.21, −0.92], Cohen's $d = -0.52$. When decomposing the three-way interaction in terms of the two-way interaction of PEs × condition for the

immediate post-test and the delayed post-test, neither of the two-way interactions was significant (Immediate Post-test: $B = -1.87$, $SE = 1.51$, $t(57) = -1.25$, $p = 0.22$, 95% CI [−4.89, 1.14] $\eta_p^2 = 0.03$; Delayed Post-test: $B = 1.62$, $SE = 1.67$, $t(57) = 0.98$, $p = 0.33$, 95% CI [−1.71, 4.96], $\eta_p^2 = 0.02$. This is no surprise, as there was no overall PEs × condition effect. However, when looking at the beta-values for the 2-way interaction, their opposite algebraic sign is noticeable, showing a cross-over. As such, the three-way interaction is a result of this cross-over effect pattern.

In the immediate post-test (see **Figure 5**), the learning conditions did not differ for lower levels (−1*SD*) of initial PEs, $MD = 2.38$, $SE = 2.14$, $p = 0.27$, 95% CI [−1.89, 6.66], Cohen's $d = 0.29$ but did so for higher levels (+1*SD*), $MD = 6.13$, $SE = 2.07$, $p = 0.004$, 95% CI [1.98, 10.27], Cohen's $d = 0.77$. As such, problem-solving was beneficial for pupils with higher initial PEs in the immediate post-test.

This pattern reverses for the delayed post-test (see **Figure 6**): For lower initial PEs, problem-solving outperformed worked-examples, $MD = 5.15$, $SE = 2.37$, $p = 0.034$, 95% CI [0.41, 9.89], Cohen's $d = 0.56$, but there was no difference for higher levels, $MD = 1.90$, $SE = 2.30$, $p = 0.41$, 95% CI [−2.68, 6.48], Cohen's $d = 0.22$.

Now we will look at the change in post-test performance over time (see **Figure 7**). Those with lower initial PEs in the worked-example condition showed a significant performance decline, $MD = -5.53$, $SE = 1.74$, $p = 0.002$, 95% CI [−9.19, −2.04], Cohen's $d = -0.55$, as did those in the problem-solving condition, $MD = -2.76$, $SE = 1.27$, $p = 0.034$, 95% CI [−5.31, −0.21], Cohen's $d = -0.39$. Although the performance decline in the problem-solving condition appears less pronounced, statistically both are comparable, $MD = -2.77$, $SE = 2.16$, $p = 0.21$, 95% CI [−1.55, 7.09], Cohens' $d = 0.33$.

For higher levels of initial PEs, those in the worked-example condition showed a comparable performance, $MD = -0.86$, $SE = 1.35$, $p = 0.53$, 95% CI [−1.84, 3.56], Cohen's $d = 0.11$, while the performance declined in the problem-solving condition, $MD = -3.37$, $SE = 1.59$, $p = 0.039$, 95% CI [−6.57, −0.17], Cohen's $d = -0.38$. These slopes in performance change were statistically significant, $MD = -4.23$, $SE = 2.01$, $p = 0.047$, 95% CI [−8.41, −0.05], Cohens' $d = -0.53$.

The Johnson-Neyman region of significance for the moderator (PROCESS, Hayes, 2018): PEs had a significant effect on changes in performance scores across both post-tests for all pupils with a (standardized) PE score of greater than 0.96.

These findings can be interpreted in the following way: For pupils with higher PEs, problem-solving in contrast to worked-examples was more beneficial resulting in an initial performance advantage. However, this early performance advantage could not be maintained in the delayed test (that is, 3 months later). The decline in performance represents the greater performance losses for higher PEs in the problem-solving condition in contrast to the worked-example condition, where performance across time was stable.

For those with lower PEs, immediate performance was not enhanced differently from either learning condition, but pupils who had learned with problem-solving showed higher delayed test scores than pupils who had learned with worked-examples. Descriptively, this is due to less pronounced performance declines over time for problem-solving in contrast to worked-examples, although the rates of performance decline are statistically not different.

Later Performance Expectancies Over Time as Metacognitive Assessments

We argued that problem-solving may influence the resulting metacognitive PEs after learning and testing by reducing overconfidence, predicting an interaction of time × condition (H4). For the analyses (see **Figure 8**), we averaged the PEs (in points) measured after presenting the example test problem of type 1 and its solution (PE1-PS1, PE2-PS1). We summed up this value with the PEs of example test problem type 2 (PE1-PS2). The resulting value represents the PEs in points (from 0 to 42) *before* both problem types had been worked on differently due to the experimental conditions (that is, PEs *prior* to learning).[6] We also summed up the PEs *after* learning with problem-type 1 and problem-type 2 (PE3-PS1, PE2-PS2; that is, PEs after learning). The same applies to the sum score of the post-test PEs for problem set 1 and problem set 2 *after* the first performance post-test (PE4-PS1, PE3-PS2 – that is, PEs after the immediate post-test at Session 1). At Session 2 and *prior* to the delayed performance test, the PEs for both problem types were summed up as well (PE5-PS1, PE4-PS2; that is, PEs prior the delayed post-test).

We subjected these indices of PEs to a repeated measures analysis of variance with condition as between-subject factor (2 levels: 0 = worked-example, 1 = problem-solving) and PEs (4 levels: prior to learning, after learning, after Test 1, prior Test 2) as within-subject variable (see **Figure 8**). Since the sphericity assumption was not met, we report the Greenhouse-Geisser-corrected p-values and degrees of freedom. This yielded a significant effect of time, $F(2.28, 129.91) = 13.41$, $p < 0.001$, $\eta_p^2 = 0.19$, but neither offered convincing evidence for a condition

[6]The results are similar when using merely the sum of PE1-PS1 and PE1-PS2 as initial performance expectancies.

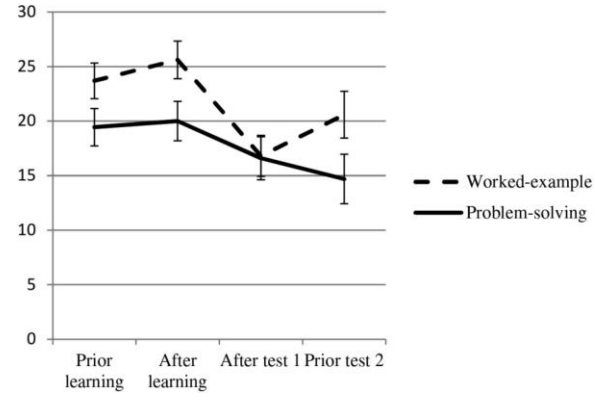


FIGURE 8 | Change in performance expectancies by learning conditions. Error bars represent the standard error of the mean. Values could range from expecting 0 points to 42 points in the performance tests.

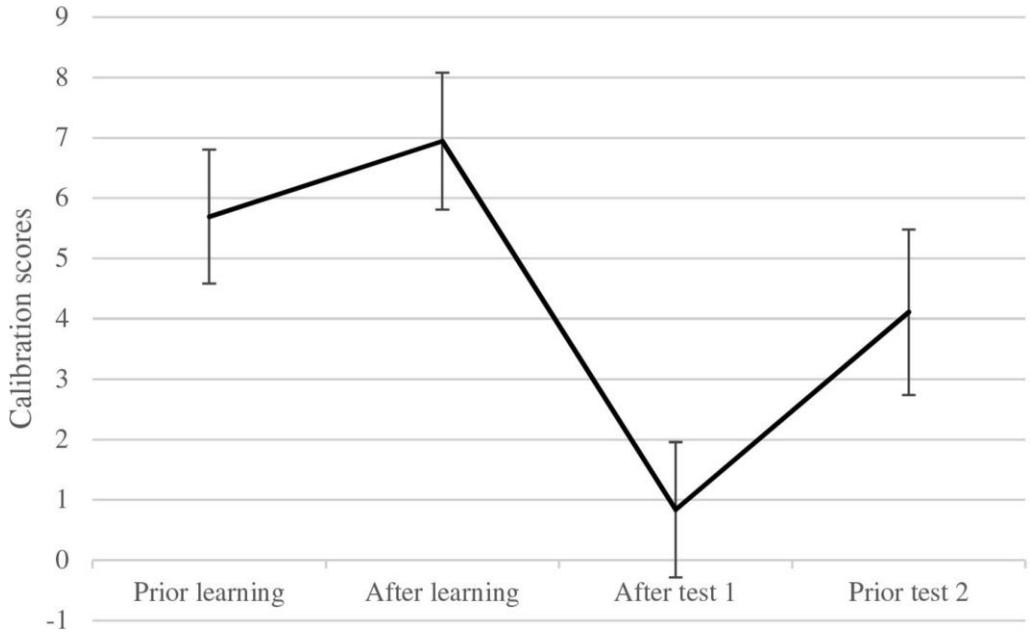


FIGURE 9 | Main effect of calibration accuracy. Error bars represent the standard error of the mean.

effect, $F(1,57) = 3.02$, $p = 0.088$, $\eta_p^2 = 0.05$, nor for the predicted interaction effect (H4), $F(2.28, 129.91) = 2.61$, $p = 0.07$, $\eta_p^2 = 0.04$.

We found little convincing support for H4. A reduction of PEs (and learner's competence illusion) was only apparent as gradual change across assessment times (see **Figure 8**); with pre-existing differences (albeit non-significant) in the worked-example condition compared to the problem-solving condition. Thus, these results should be taken with caution.

Calibration (Metacognitive Accuracy)

To obtain calibration (difference of predicted and actual test scores), we used the PEs (previously discussed in **Figure 8**) and the actual test scores: We computed a difference score of PEs prior to learning and immediate post-test performance; a difference score of PEs after learning and immediate post-test performance; a difference score of PEs after the immediate post-test and actual test performance in the immediate post-test; and, a difference score of later PEs prior the delayed post-test and actual performance in the delayed post-test. (Note, positive values denote overconfidence and negative ones underconfidence; Bugg and McDaniel, 2012).

Using these calibration values as dependent variables (within-subjects; 4 levels: calibration prior to learning, calibration after learning, calibration after the immediate post-test, calibration prior to the delayed post-test) and condition as independent variable (between-subjects) in an rANOVA yielded a main effect of condition, $F(1,57) = 12.32$, $MD = -6.78$, $SE = 1.93$, $p < 0.001$, 95% CI [−10.65, −2.91], $\eta_p^2 = 0.18$. Pupils in the worked-example group showed less accurate calibration and more overconfidence, $M = 7.79$, $SE = 1.33$, 95% CI [5.12, 10.45], while pupils' calibrations in the problem-solving group was more accurate, $M = 1.01$, $SE = 1.40$, 95% CI [−1.80, 3.81]. Note that the calibration score of the problem-solving group is closer to 0, which denotes more accurate calibration, whereas a score of 7.79 in the worked-example group represents a difference of about 8 points between expectation and actual test scores.

We further found a main effect of calibration (reported with Greenhouse-Geisser correction), $F(2.32,132.45) = 10.84$, $p < 0.001$, $\eta_p^2 = 0.16$ (see **Figure 9**). Simple comparisons (Bonferroni-corrected) showed a significant difference of calibration *prior* to learning and calibration *after* the immediate post-test ($M = 0.83$, $SE = 1.12$), $MD = 4.86$, $SE = 1.04$, $p < 0.001$, 95% CI [2.01, 7.71] and a significant difference of calibration *after* learning and calibration *after* the immediate post-test, $MD = 6.11$, $SE = 1.06$, $p < 0.001$, 95% CI [3.20, 9.01]. This means calibration *after* the immediate post-test was more accurate than *prior* to and after learning. All other comparisons were not significant, all ps > 0.15. Finally, we did not find the expected interaction effect of calibration × condition, $F(2.32,132.43) = 2.17$, $p = 0.11$, 95% CI [2.99, 10.73], $\eta_p^2 = 0.04$. Overall this pattern indicates that the calibration in the problem-solving condition was more accurate than in the worked-example condition in general (but not as a consequence of the learning conditions or tests over time), and that calibration after the immediate test was more accurate than PEs prior to both tests. This pattern of results partially supports (H5). Overall calibration in the problem-solving condition was more accurate as in the worked-example condition showing overconfidence. However, due to the pre-existing differences (albeit non-significant) in initial PEs in the worked-example condition compared to the problem-solving condition (see **Figure 8**), the interpretation of the results on calibration due to overconfidence reduction

is not routed in strong empirical evidence and should be taken with caution.

DISCUSSION

Our work examined learners' PEs prior to learning as moderators for the effectiveness of different learning tasks (a special type of problem-solving vs. worked-examples) on immediate and delayed performance. The experiment was conducted in school and used curricular mathematical materials for learning. We assumed that the problem-solving condition would be superior to the worked-examples condition in the delayed post-test (time × condition; H1) and that problem-solving opposed to worked-examples are more beneficial for higher PEs (condition × PEs; H2). We further supposed that the moderating effect of PEs in the problem-solving condition would arise particularly in the delayed test (time × condition × PEs; H3). Moreover, we predicted an interaction effect of condition and time on metacognitive judgments of PEs measured after learning and testing (H4). Participants in the problem-solving condition (in comparison to participants in the worked-examples condition) should lower their PEs regarding the later test outcome after experiencing the difficult learning task (reduction of competence illusion). Finally, we also assumed that calibration accuracy (the difference between expected performance and actual performance) should be more precise for problem-solvers in contrast to participants in the worked-examples condition (H5). Participants in the worked-example condition probably maintain a competence misconception and thereby may have stronger differences between their expected and their actual performance. Thus, we expected initial PEs to be a moderator for learning performance and condition to be a moderator for later PEs, thus affecting metacognitive accuracy.

Our findings showed only a descriptive advantage of the problem-solving condition on the delayed learning performance (H1) and no two-way interaction of PEs and the condition (H2). However, taking into account prior PEs, we obtained a beneficial adjusted main effect of the problem-solving condition for participants with average PEs. Thus, problem-solving can be advantageous for certain learners. This is in line with the assumptions that PEs are only related to difficult (and not easy) tasks (like problem-solving) because difficult tasks require more effort, time, motivation, and persistence (e.g., Marshall and Brown, 2004; Reinhard and Dickhäuser, 2009). The obtained moderation supports the notion that learner characteristics are important for the effectiveness of desirable difficulties (e.g., McDaniel and Butler, 2011). For pupils with lower and average PEs, the problem-solving condition was more advantageous later on, while for higher PEs both learning conditions were equal at a delay. This is partly in line with the assumptions that the beneficial effects of generation tasks arise in the long run (e.g., Bjork and Bjork, 1992, 2011; Bjork, 1994): There was no significant interaction between time and condition, and only the consideration of initial PEs unveiled favorable effects at a delay. Without taking into account PEs, performance in the problem-solving condition was only descriptively better long-term; this could be due to the long delay between learning and the delayed test (this will be further discussed below).

The three-way interaction (PEs × time × condition) showed that participants with higher PEs in the problem-solving condition performed better in the immediate test, whereas participants with lower PEs in the problem-solving condition performed better in the delayed test. Unfortunately, higher PEs could not maintain this initial performance advantage in the problem-solving condition over time. Although participants with higher expectancies immediately profited from generation tasks, those with lower PEs also benefited from difficult tasks in the long run. Thus, as inquired in the beginning, it is not strange to trouble a learner who has lower PEs with hindered learning tasks. This is in line with the assumptions that desirable difficulties may be advantageous for learners with lower abilities or cognitive motivation (e.g., McDaniel et al., 2002; Schindler et al., 2019). It is important to note that these difficulties do not boost weaker learners' performances to the level of stronger learners, but these difficulties prevent greater performance losses for weaker learners over time.

Overall, learners benefited in different manners from desirable difficulties. This fits previous work that was able to identify moderators (e.g., feedback, mood, previous knowledge, reading skills; e.g., McNamara et al., 1996; McDaniel et al., 2002; Bertsch et al., 2007; Schindler et al., 2017). The present findings also emphasize the importance of moderators for the effectiveness of generation activities.

When considering the effects of a generation activity on metacognitions, the results have to be taken with caution. A mere trend shows a gradual decrease in PEs in the problem-solving condition in contrast to the worked-examples condition over time (in which overconfident PEs did not change; H4); and a trend shows a pre-existing difference in PEs. The results show no convincing support for a learning event and time-driven overconfidence reduction (H4). Regarding our fifth hypotheses, our results showed a main effect of condition with greater metacognitive accuracy in the problem-solving condition than in the worked-examples condition (H5). Thus, calibration accuracy (the difference between expected performance and actual performance) was more precise for participants in the problem-solving condition in contrast to participants in the worked-examples condition. Yet, this interpretation is not routed in strong empirical evidence and should be taken with caution. These findings only hint that the problem-solving task may have led to a more realistic understanding of learners' current competences and thus reduced participants' competence illusion (e.g., Karpicke et al., 2009; Diemand-Yauman et al., 2011; Baars et al., 2014). Given the important role of accurate metacognitions for the regulation of learning (e.g., Dunlosky and Lipko, 2007), these findings nevertheless hint at the value of problem-solving.

The current study is not without some limitations, which will be discussed in the following section and which could be optimized in future work. We designed our study with real-world materials that were integrated within curricular content and natural math lesson progression. Although we coordinated with the teachers on what content was covered prior to our experimental session (introduction of the topic but no

calculations), we had no control over actual implementations (although there were no differences in previous knowledge across both conditions). Moreover, after the first experimental in-class session, we had no control over any further progression of the lessons' content, over homework or over subsequent topics, prior to the delayed test of Session 2. The teachers knew about the delayed test and were instructed not to repeat any content; however, we do not know what additional content with potential overlap was taught in the interim between Session 1 and Session 2, and/or what pupils learned in the meantime. Thus, although classroom studies are very important regarding work focusing on learning success, there are also many confounding factors that are not controllable, which presents a limitation. Performance in general was rather low, thus it would be interesting to extend the instructional units.

Another limitation relates to the fact that the tests in Sessions 1 and 2 were identical, thus the testing-effect may have played a role regarding learners' performance, although likely not much given the 3-month delay. To avoid this, future studies may include one group tested immediately and another group only tested at a delay. In addition, our worked-examples included detailed explanations, so it may be that learners did not have to indulge in self-explanations (which can trigger the positive effects of worked-examples; e.g., Renkl et al., 1998). Hence, future research could use materials that require self-explanations. In line with this, it could also be that our problem-solving condition was superior to the worked-example condition not because of the generation task but because participants were presented with a shortened worked-example before the generation activity (see e.g., Paas, 1992), as well as briefly with the correct answers after the generation, and feedback is important for the effectivity (e.g., Slamecka and Fevreiski, 1983; Pashler et al., 2005; Kang et al., 2007; Metcalfe and Kornell, 2007; Potts and Shanks, 2014; Metcalfe, 2017). Thus, future studies could use different incantations of problem-solving tasks or worked-examples, all in the attempt to generalize findings and to try to optimize possible limitations due to our applied learning tasks. In line with this, in the applied problem-solving tasks students were able to look back at the explanations and introduction of the material given in the beginning (open book solving task). Although this, as well as later given feedback sheet regarding correct answers for the generation tasks, may have been beneficial, it is unclear to what extent students even used these aids. Some students may have never looked at the previous learning materials, whereas others may have relied on them often; some may have contemplated the correct solution steps after finding out a discrepancy in their results and the result provided on the answer sheet, others may have not. Although this is a typical occurrence in schools, future work could also try to manipulate how many times learners are able to look back at previously studied materials. Previous work also often implemented problem-solving tasks after worked-examples, thus combining these two strategies. In contrast, we compared sole problem-solving tasks and sole studying of worked-examples (both following a short introduction of the materials), so our methods are not completely in line with some of the above-mentioned literature. Future studies could thus explore the relation of PEs, problem solving following worked-examples, and long-time learning success.

A further, and possibly confounding or negative, aspect concerns the lag between post-test one and post-test two, which we set at 3 months. The 3-month lag taps into long-term learning but may have been too long given the overall low performance. Future research may include a shorter lag of only a few weeks. However, the choice of 3 months was implemented because we wanted to make sure that all teachers had finished the section on linear functions; naturally, the length of time dedicated to a topic depends on the teachers and on the class (in other words, some classes progress more quickly than others), which we cannot influence due to the field character of our study. In our case, we aimed for a comparable lag and for all teachers to have started new content so that the end-of-topic exam on linear functions did not coincidentally occur in temporal proximity to our delayed test. It would be valuable for future research to coordinate with teachers' planned exam at the end of the session to include mutually agreed-upon exam questions that would also serve as a delayed test. One related problem/aspect of that strategy (and our research) would be that any previous one-time intervention may be too weak to detect differences in delayed exam performance as it may be overshadowed by teachers' and students' own exam preparations (which we cannot control). Relatedly, a single-intervention study may have to be paired with a shorter lag, or multiple controlled interventions are required for longer lags. The difficulty here lies in the willingness of the teachers and parents to participate, given real-world constraints and concerns that these interventions could disrupt the classes and take away valuable teaching time. Future research may also conceptualize a paradigm in which trained teachers take over teaching for one to 3 weeks, with multiple, ongoing experimental interventions that conclude with a graded interim exam as a delayed test. This may present the challenge of finding willing institutions, teachers, and/or parents.

To thoroughly test moderators, larger samples are needed (which is often difficult to obtain in school contexts). Of course, our findings can be interpreted only for German students within the same age-range, the same educational school track, and for the same learning materials (and very strictly seen, only for this school). Due to that, future work using bigger and more diverse samples (as well as different materials) is important. The same applies to learners with different levels of previous knowledge: Future studies could use more known topics, assess previous knowledge, and include this factor in the analyses. To gain access to more participants, another option for future research may include extracurricular learning environments (e.g., instead of homework), which could be implemented either online or onsite. For instance, a study could deploy carefully designed learning modules on selected (additional) curricular content that is not part of current class curriculum within a given school year; this might allow the implementation of thorough experimental designs while proving attractive to learners and teachers as a supplemental training learning environment. All in all, as pointed out by Dunlosky et al. (2013), future research

may attend in general more to an investigation of moderators of various desirable difficulties (e.g., previous knowledge, different skill levels) because their roles are still less known.

We should note that previous work often focused on the effectiveness of generation tasks regarding recall and/or memory of learned information through later tests assessing the same or similar information, but our tests mostly assessed transfer (instead of identical information). Thus, the underlying effects of the learning conditions could be different (e.g., Glogger-Frey et al., 2015). Prior research regarding transfer and intentionally aggravated learning tasks resulted in varying findings: Some studies found beneficial effects of desirable difficulties solely for identical or easy information but not for transfer (e.g., Lehmann et al., 2016) or that worked examples were more important for transfer (e.g., Glogger-Frey et al., 2015). In contrast, some studies found beneficial effects of desirable difficulties also for changed materials and transfer (see e.g., Dunlosky et al., 2013 for a good overview). Thus, future studies could implement transfer as well as identical questions.

As mentioned above, generation tasks reduce learners' competence illusion and overconfidence, thus participants in the problem-solving condition should be able to more accurately calibrate their PEs than do participants in the worked-example condition, who could still possess overconfident expectancies. Our findings only hint at this relationship. Participants' PEs appeared to differ between the conditions before the learning tasks even started. This does not have to be an indicator that the randomization of our sample failed but could rather indicate that participants (unbeknown be us) checked the tasks and their condition by looking at the materials prior to the learning task, which serves as a limitation. Hence, their initial PEs could have been influenced by participants' knowledge of the upcoming learning tasks.

CONCLUSION

Our results emphasize the importance of moderators for the desirability of generation activities, and the desirability of generation activities for metacognitive outcomes. Regarding implications for the educational context, we still cannot recommend that teachers use or not use problem-solving tasks. Our work, though, is a step in the right direction, while more research exploring the effectiveness of problem-solving tasks or moderators are still needed. Thus, we underscore the value of longitudinal studies or studies using multiple learning phases as well as multiple learning success assessments for evidence-driven educational recommendations.

ETHICS STATEMENT

This study was conducted in full accordance with the Ethical Guidelines of the German Association of Psychologists (DGPs), the American Psychological Association (APA), and the Hessian Ministry for Science and the Arts. The study was approved by the Hessian Ministry for Science and the Arts. Full consent was obtained of the principal, the teachers, parents and pupils.

AUTHOR CONTRIBUTIONS

M-AR provided funding, developed the basic idea on performance expectancies as moderator, and provided the critical comments. SW developed the basic idea of performance expectancies as metacognitions, developed the materials, contributed to the data collection and data preparation. M-AR and SW contributed to the study design and analyzed the data. SW and KW wrote and revised the manuscript.

FUNDING

This study was supported by the Federal State of Hessen and its LOEWE research initiative Desirable Difficulties in Learning [LOEWE: Landes-Offensive zur Entwicklung wissenschaftlich-ökonomischer Exzellenz (state offensive for the development of scientific and economic excellence)], project "desirable difficulties; intrinsic cognitive motivation and performance expectancies" awarded to M-AR.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01623/full#supplementary-material

REFERENCES

Adesope, O. O., Trevisan, D. A., and Sundararajan, N. (2017). Rethinking the use of tests: a meta-analysis of practice testing. *Rev. Educ. Res.* 87, 659–701. doi: 10.3102/0034654316689306

Agarwal, P. K., Karpicke, J. D., Kang, S. H., Roediger, H. L., and McDermott, K. B. (2008). Examining the testing effect with open-and closed-book tests. *Appl. Cogn. Psychol.* 22, 861–876. doi: 10.1002/acp.1391

Ajzen, I., and Fishbein, M. (1988). *Theory of Reasoned Action-Theory of Planned Behavior*. Tampa, FL: University of South Florida.

Baars, M., Gog, T., Bruin, A., and Paas, F. (2014). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Appl. Cogn. Psychol.* 28, 382–391. doi: 10.1002/acp.3008

Baars, M., Van Gog, T., de Bruin, A., and Paas, F. (2016). Effects of problem solving after worked example study on secondary school children's monitoring accuracy. *Educ. Psychol.* 37, 810–834. doi: 10.1080/01443410.2016.1150419

Berthold, K., and Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. *J. Educ. Psychol.* 101, 70–87. doi: 10.1037/a0013247

Bertsch, S., Pesta, B. J., Wiscott, R., and McDaniel, M. A. (2007). The generation effect: a meta-analytic review. *Mem. Cogn.* 35, 201–210. doi: 10.3758/bf03193441

Bjork, E. L., and Bjork, R. A. (2011). "Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning," in *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, eds M. A.

Gernsbacher, R. W. Pew, L. M. Hough, and J. R. Pomerantzcpesnm, (New York, NY: Worth Publishers), 59–68.

Bjork, R. A. (1994). "Memory and metamemory considerations in the training of human beings," in *Metacognition: Knowing about Knowing*, eds J. Metcalfecpesnm, and A. Shimamuracpesnm, (Cambridge, MA: MIT Press), 185–205.

Bjork, R. A., and Bjork, E. L. (1992). "A new theory of disuse and an old theory of stimulus fluctuation," in *From learning processes to cognitive processes: Essays in honor of William K. Estes, Volume* II, eds A. F. Healy, S. M. Kosslyn, and R. M. Shiffrincpesnm, (London: Psychology Press), 35–67.

Brewer, G. A., and Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *J. Mem. Lang.* 66, 407–415. doi: 10.1016/j.jml.2011.12.009

Bugg, J. M., and McDaniel, M. A. (2012). Selective benefits of question self-generation and answering for remembering expository text. *J. Educ. Psychol.* 104, 922–931. doi: 10.1037/a0028661

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., and Rohrer, D. (2006). Distributed practice in verbal recall tasks: a review and quantitative synthesis. *Psychol. Bull.* 132, 354–380. doi: 10.1037/0033-2909.132.3.354

Crouch, C., Fagen, A. P., Callan, J. P., and Mazur, E. (2004). Classroom demonstrations: learning tools or entertainment? *Am. J. Phys.* 72, 835–838. doi: 10.1119/1.1707018

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13, 319–340.

DeWinstanley, P. A., and Bjork, E. L. (2004). Processing strategies and the generation effect: implications for making a better reader. *Mem. Cogn.* 32, 945–955. doi: 10.3758/bf03196872

Dickhäuser, O., and Reinhard, M. A. (2006). Factors underlying expectancies of success and achievement: the influential roles of need for cognition and general or specific self-concepts. *J. Pers. Soc. Psychol* 90, 490–500. doi: 10.1037/0022-3514.90.3.490

Diemand-Yauman, C., Oppenheimer, D. M., and Vaughan, E. B. (2011). Fortune favors the (): Effects of disfluency on educational outcomes. *Cognition* 118, 111–115. doi: 10.1016/j.cognition.2010.09.012

Dunlosky, J., and Lipko, A. R. (2007). Metacomprehension: a brief history and how to improve its accuracy. *Curr. Dir. Psychol. Sci.* 16, 228–232. doi: 10.1111/j.1467-8721.2007.00509.x

Dunlosky, J., and Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: Sage Publications.

Dunlosky, J., and Rawson, K. A. (2012). Overconfidence produces underachievement: inaccurate self evaluations undermine students' learning and retention. *Learn. Instruct.* 22, 271–280. doi: 10.1016/j.learninstruc.2011.08.003

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., and Willingham, D. T. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychol. Sci. Public Interest* 14, 4–58. doi: 10.1177/1529100612453266

Eagly, A. H., and Chaiken, S. (1993). *The Psychology of Attitudes*. San Diego, CA: Harcourt Brace Jovanovich College Publishers.

Ebbinghaus, H. (1913). *Memory: A Contribution to Experimental Psychology*. Trans. H. A. Ruger and C. E. Bussenius. (New York, NY: Teachers College Press).

Eccles, J. S. (1983). "Expectancies, values, and academic behaviors," in *Achievement and achievement motives: Psychological and sociological approaches*, ed. J. T. Spencecpesnm, (San Francisco, CA: Free man), 75–146.

Eccles, J. S., and Wigfield, A. (2002). Motivational beliefs, values, and goals. *Ann. Rev. Psychol.* 53, 109–132. doi: 10.1146/annurev.psych.53.100901.135153

Fiorella, L., and Mayer, R. E. (2016). Eight ways to promote generative learning. *Educ. Psychol. Rev.* 28, 717–741. doi: 10.1007/s10648-015-9348-9

Glogger-Frey, I., Fleischer, C., Grüny, L., Kappich, J., and Renkl, A. (2015). Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. *Learn. Instruct.* 39, 72–87. doi: 10.1016/j.learninstruc.2015.05.001

Griffin, T. D., Jee, B. D., and Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Mem. Cogn.* 37, 1001–1013. doi: 10.3758/MC.37.7.1001

Hayes, A. F. (2018). *Introduction to Mediation, Moderation, and Conditional Process Analysis*, 2nd Edn. New York, NY: The Guilford Press.

Hefter, M. H., Berthold, K., Renkl, A., Riess, W., Schmid, S., and Fries, S. (2014). Effects of a training intervention to foster argumentation skills while processing conflicting scientific positions. *Instr. Sci.* 42, 929–947. doi: 10.1007/s11251-014-9320-y

Judd, C. M., Kenny, D. A., and McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychol. Methods* 6, 115–134. doi: 10.1037//1082-989x.6.2.115

Kalyuga, S., Ayres, P., Chandler, P., and Sweller, J. (2003). The expertise reversal effect. *Educ. Psychol.* 38, 23–31. doi: 10.1207/s15326985ep3801_4

Kalyuga, S., Chandler, P., Tuovinen, J., and Sweller, J. (2001). When problem solving is superior to studying worked examples. *J. Educ. Psychol.* 93, 579–588. doi: 10.1037/0022-0663.93.3.579

Kalyuga, S., and Renkl, A. (2010). Expertise reversal effect and its instructional implications: introduction to the special issue. *Instr. Sci.* 38, 209–215. doi: 10.1007/s11251-009-9102-0

Kang, S. H., McDermott, K. B., and Roediger, H. L. III (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *Eur. J. Cogn. Psychol.* 19, 528–558. doi: 10.1080/09541440601056620

Kant, J. M., Scheiter, K., and Oschatz, K. (2017). How to sequence video modeling examples and inquiry tasks to foster scientific reasoning. *Learn. Instr.* 52, 46–58. doi: 10.1016/j.learninstruc.2017.04.005

Karpicke, J. D., and Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science* 331, 772–775. doi: 10.1126/science.1199327

Karpicke, J. D., Butler, A. C., and Roediger, H. L. III (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own? *Memory* 17, 471–479. doi: 10.1080/09658210802647009

Karpicke, J. D., and Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *J. Mem. Lang.* 62, 227–239. doi: 10.1016/j.jml.2009.11.010

Koondhar, M. Y., Molok, A., Nuha, N., Chandio, F., Rind, M. M., Raza, A., et al. (2015). "A conceptual framework for measuring the acceptance of pervasive learning," in *Proceedings of the 5th International Conference on Computing and Informatics*, (Turkey).

Koriat, A. (1997). Monitoring one's own knowledge during study: a cue-utilization approach to judgments of learning. *J. Exp. Psychol. Gen.* 126, 349–370. doi: 10.1037/0096-3445.126.4.349

Koriat, A., and Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *J. Exp. Psychol.* 31, 187–194. doi: 10.1037/0278-7393.31.2.187

Koriat, A., and Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Mem. Cogn.* 34, 959–972. doi: 10.3758/BF03193244

Kornell, N., Rabelo, V. C., and Klein, P. J. (2012). Tests enhance learning – Compared to what? *J. Appl. Res. Mem. Cogn.* 1, 257–259. doi: 10.1016/j.jarmac.2012.10.002

Kruglanski, A. W., and Stroebe, W. (2005). "The influence of beliefs and goals on attitudes: Issues of structure, function, and dynamics," in *The Handbook of Attitudes*, eds D. Albarracincpesnm, and B. T. Johnsoncpesnm, (Abingdon: Routledge), 323–368.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., et al. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente*. Berlin: Max-Planck-Institiut für Bildungsforschung.

Lehmann, J., Goussios, C., and Seufert, T. (2016). Working memory capacity and disfluency effect: an aptitude-treatment-interaction study. *Metacogn. Learn.* 11, 89–105. doi: 10.1007/s11409-015-9149-z

Marshall, M. A., and Brown, J. D. (2004). Expectations and realizations: The role of expectancies in achievement settings. *Motiv. Emot.* 28, 347–361. doi: 10.1007/s11031-004-2388-y

McDaniel, M. A., and Butler, A. C. (2011). "A contextual framework for understanding when difficulties are desirable," in *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork*, ed. A. S. Benjamincpesnm, (London: Psychology Press), 175–198.

McDaniel, M. A., Hines, R. J., and Guynn, M. J. (2002). When text difficulty benefits less-skilled readers. *J. Mem. Lang.* 46, 544–561. doi: 10.1006/jmla.2001.2819

McDaniel, M. A., Waddill, P. J., and Einstein, G. O. (1988). A contextual account of the generation effect: a three-factor theory. *J. Mem. Lang.* 27, 521–536. doi: 10.1016/0749-596x(88)90023-x

McNamara, D. S., Kintsch, E., Songer, N. B., and Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cogn. Instr.* 14, 1–43. doi: 10.1207/s1532690xci1401_1

Metcalfe, J. (2017). Learning from errors. *Ann. Rev. Psychol.* 68, 465–489. doi: 10.1146/annurev-psych-010416-044022

Metcalfe, J., and Kornell, N. (2007). Principles of cognitive science in education: the effects of generation, errors, and feedback. *Psychon. Bull. Rev.* 14, 225–229. doi: 10.3758/bf03194056

Mihalca, L., Mengelkamp, C., and Schnotz, W. (2017). Accuracy of metacognitive judgments as a moderator of learner control effectiveness in problem-solving tasks. *Metacogn. Learn.* 12, 357–379. doi: 10.1007/s11409-017-9173-2

Moreno, R., Reisslein, M., and Ozogul, G. (2009). Optimizing worked-example instruction in electrical engineering: the role of fading and feedback during problem-solving practice. *J. Eng. Educ.* 98, 83–92. doi: 10.1002/j.2168-9830.2009.tb01007.x

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J. Educ. Psychol.* 84, 429–434. doi: 10.1037/0022-0663.84.4.429

Pashler, H., Cepeda, N. J., Wixted, J. T., and Rohrer, D. (2005). When does feedback facilitate learning of words? *J. Exp. Psychol.* 31, 3–8. doi: 10.1037/0278-7393.31.1.3

Pieger, E., Mengelkamp, C., and Bannert, M. (2017). Fostering analytic metacognitive processes and reducing overconfidence by disfluency: the role of contrast effects. *Appl. Cogn. Psychol.* 31, 291–301. doi: 10.1002/acp.3326

Potts, R., and Shanks, D. R. (2014). The benefit of generating errors during learning. *J. Exp. Psychol. Gen.* 143, 644–667. doi: 10.1037/a0033194

Reinhard, M. A., and Dickhäuser, O. (2009). Need for cognition, task difficulty, and the formation of performance expectancies. *J. Pers. Soc. Psychol.* 96, 1062–1076. doi: 10.1037/a0014927

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cogn. Sci.* 38, 1–37. doi: 10.1111/cogs.12086

Renkl, A., and Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: a cognitive load perspective. *Educ. Psychol.* 38, 15–22. doi: 10.1207/s15326985ep3801_3

Renkl, A., Atkinson, R. K., Maier, U. H., and Staley, R. (2002). From example study to problem solving: smooth transitions help learning. *J. Exp. Educ.* 70, 293–315. doi: 10.1080/00220970209599510

Renkl, A., Stark, R., Gruber, H., and Mandl, H. (1998). Learning from worked-out examples: the effects of example variability and elicited self-explanations. *Contemp. Educ. Psychol.* 23, 90–108. doi: 10.1006/ceps.1997.0959

Richland, L. E., Bjork, R. A., Finley, J. R., and Linn, M. C. (2005). "Linking cognitive science to education: Generation and interleaving effects," in *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*, (Mahwah, NJ: Erlbaum).

Roediger, H. L. III, and Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychol. Sci.* 17, 249–255. doi: 10.1111/j.1467-9280.2006.01693.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychol. Bull.* 140, 1432–1463. doi: 10.1037/a0037559

Schindler, J., Richter, T., and Eyßer, C. (2017). Mood moderates the effect of self-generation during learning. *Front. Learn. Res.* 5, 76–88. doi: 10.14786/flr.v5i4.296

Schindler, J., Schindler, S., and Reinhard, M.-A. (2019). Effectiveness of self-generation during learning is dependent on individual differences in need for cognition. *Front. Learn. Res.* 7, 23–39. doi: 10.14786/flr.v7i2.407

Schindler, S., Reinhard, M. A., and Dickhäuser, O. (2016). Boon and bane of being sure: the effect of performance certainty and expectancy on task performance. *Eur. J. Psychol. Educ.* 31, 245–253. doi: 10.1007/s10212-015-0267-4

Schworm, S., and Renkl, A. (2006). Computer-supported example-based learning: when instructional explanations reduce self-explanations. *Comput. Educ.* 46, 426–445. doi: 10.1016/j.compedu.2004.08.011

Serra, M. J., and Metcalfe, J. (2009). "15 Effective Implementation of Metacognition," in *Handbook of Metacognition in Education*, eds A. C. Graesser, D. J. Hacker, and J. Dunloskycpesnm, (Abingdon: Routledge), 278.

Slamecka, N. J., and Fevreiski, J. (1983). The generation effect when generation fails. *J. Verb. Learn. Verb. Behav.* 22, 153–163. doi: 10.1016/s0022-5371(83)90112-3

Son, L. K., and Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *J. Exp. Psychol.* 26, 204–221. doi: 10.1037//0278-7393.26.1.204

Spanjers, I. A., Wouters, P., Van Gog, T., and Van Merrienboer, J. J. (2011). An expertise reversal effect of segmentation in learning from animated worked-out examples. *Comput. Hum. Behav.* 27, 46–52. doi: 10.1016/j.chb.2010.05.011

Sweller, J. (2006). The worked example effect and human cognition. *Learn. Instr.* 16, 165–169. doi: 10.1016/j.learninstruc.2006.02.005

Thiede, K. W., Anderson, M., and Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *J. Educ. Psychol.* 95, 66–73. doi: 10.1348/135910710X510494

van Gog, T., and Kester, L. (2012). A test of the testing effect: acquiring problem-solving skills from worked examples. *Cogn. Sci.* 36, 1532–1541. doi: 10.1111/cogs.12002

van Gog, T., Kester, L., Dirkx, K., Hoogerheide, V., Boerboom, J., and Verkoeijen, P. (2015). Testing after worked example study does not enhance delayed problem-solving performance compared to restudy. *Educ. Psychol. Rev.* 27, 265–289. doi: 10.1007/s10648-015-9297-3

Weissgerber, S. C., Reinhard, M. A., and Schindler, S. (2017). Learning the hard way: need for cognition influences attitudes toward and self-reported use of desirable difficulties. *Educ. Psychol.* 38, 176–202. doi: 10.1080/01443410.2017.1387644

Wittrock, M. C. (1989). Generative processes of comprehension. *Educ. Psychol.* 24, 345–376. doi: 10.1207/s15326985ep2404_2

Wittwer, J., and Renkl, A. (2010). How effective are instructional explanations in example-based learning? A meta-analytic review. *Educ. Psychol. Rev.* 22, 393–409. doi: 10.1007/s10648-010-9136-5

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *Am. Educ. Res. J.* 45, 166–183. doi: 10.3102/0002831207312909

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX C

Wenzel, K., & Reinhard, M.-A. (2019). Relatively unintelligent individuals do not benefit from intentionally hindered learning: The role of desirable difficulties. *Intelligence, 77,* 101405. https://doi.org/10.1016/j.intell.2019.101405

This is the final article version published by Elsevier in *Intelligence* available online: https://www.sciencedirect.com/science/article/abs/pii/S0160289619301874

Intelligence 77 (2019) 101405

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell

Relatively unintelligent individuals do not benefit from intentionally hindered learning: The role of desirable difficulties

Kristin Wenzel*, Marc-André Reinhard

Department of Psychology, University of Kassel, Holländische Straße 36-38, 34127 Kassel, Germany

ARTICLE INFO

Keywords:
Intelligence
Testing effect
Retrieval
Generation effect
Desirable difficulties
Long-term learning

ABSTRACT

Intelligence is an important predictor of long-term learning and academic achievement. In two studies we focused on the relation among intelligence, desirable difficulties–active generation/production of information and taking tests–, and long-term learning. We hypothesized that intelligence is positively correlated to long-term learning and that difficult learning situations, as opposed to easier reading, increase later long-term learning. We further assumed that the beneficial effects of difficult learning would be moderated by intelligence, thus, we supposed the positive effects to be stronger for learners with higher intelligence and weaker for learners with lower intelligence. We in turn conducted two experiments (N_1 = 149, N_2 = 176, respectively), measured participants' intelligence, applied desirable difficulties–generation/testing–in contrast to control tasks, and later assessed long-term learning indicated by delayed final test performance. Both studies showed positive correlations between intelligence and later long-term learning. Study 2 further found the expected beneficial effect of difficult learning, which was also moderated by intelligence. There was no difference between difficult tasks and control tasks for participants with relatively low intelligence. Retrieving answers in learning tests was, however, beneficial for participants with average intelligence and even more beneficial for participants with higher intelligence. In general, our two experiments highlight the importance of intelligence for complex and challenging learning tasks that are supposed to stimulate deeper encoding and more cognitive processing. Thus, specifically learners with higher, or at least average, intelligence should be confronted with difficulties to increase long-term learning and test performance.

1. Introduction

Intelligence is often defined as a mental ability that includes information processing, understanding of complex ideas, logical or analytical reasoning, problem solving, remembering information, acquiring knowledge and skills, efficiently dealing with novel tasks, and an ability to learn (e.g., Snyderman & Rothman, 1987; Stern, 2015; Sternberg, 1997). A variety of theories and paradigms explaining intelligence highlight the importance of a general mental ability or a general intelligence factor (*g*) for such cognitive processes and learning (e.g., Jäger, 1982; Spearman, 1904, 1939). Intelligence has repeatedly been shown to impact almost all aspects of daily life and was able to predict a broad array of successful human behaviors, performances, and outcomes, including increased creativity, elevated potential, improved health, better job performance, higher income, and longer employment (e.g., Fergusson, Horwood, & Ridder, 2005; Kuncel, Hezlett, & Ones, 2004; Strenze, 2007). Intelligence is further linked to an ability to solve complex problems, and is especially valuable regarding complex tasks, complex information processing, or complex situations (e.g., Gottfredson, 1997; Kuncel et al., 2004; Roth et al., 2015; Stadler, Becker, Gödker, Leutner, & Greiff, 2015). Furthermore, many studies found intelligence to be strongly and positively correlated to long-term learning and academic achievement, and it is often cited as one of the strongest predictors of long-term learning and academic achievement. This includes varying measures of learning outcomes in laboratories and classrooms, like final test performance or memory outcomes, as well as varying measures of academic success like grades, gained school qualifications, or probabilities of gaining university degrees (e.g., Bornstein, Hahn, & Wolke, 2013; Fergusson et al., 2005; Kuncel et al., 2004; Stern, 2015; Strenze, 2007; Roth et al., 2015; for an overview of multiple meta-analyses regarding intelligence and different operationalizations of success, including long-term learning and academic achievement, see Strenze, 2015).

For instance, Fergusson et al. (2005) showed in a longitudinal study spanning 25 years that children's intelligence assessed in middle childhood significantly predicted individuals' later probability of gaining school qualifications, post-school educational/vocational qualifications, and university degrees. Higher intelligence was thus consistently linked to higher academic achievement, even controlling for

* Corresponding author.
E-mail addresses: kristin.wenzel@uni-kassel.de (K. Wenzel), reinhard@psychologie.uni-kassel.de (M.-A. Reinhard).

https://doi.org/10.1016/j.intell.2019.101405
Received 3 April 2019; Received in revised form 17 September 2019; Accepted 4 October 2019

gender and early conduct problems and for family, social, and individual factors (Fergusson et al., 2005). Roth et al. (2015) conducted a meta-analysis regarding the effects of intelligence on school grades. The authors included 240 samples with 105,185 participants and found that higher intelligence was strongly linked to better school grades ($r = 0.44$, corrected population correlation: $\rho = 0.54$). They found among other conclusions that different types of intelligence tests resulted in different population correlations, the highest by mixed intelligence tests including both verbal and nonverbal tasks ($\rho = 0.60$), followed by verbal intelligence tests ($\rho = 0.53$), both significantly higher than the population correlation yielded by nonverbal tests ($\rho = 0.44$). Individuals' grade level also proved important, showing that the population correlation of intelligence and grades was significantly stronger for high school ($\rho = 0.58$) and middle school ($\rho = 0.54$) than for elementary school ($\rho = 0.45$). The authors also took a closer look at the varying subject domains and grades: The linkages between intelligence and a cluster of mathematical and science courses (e.g., mathematics, biology, and physics) were especially strong ($\rho = 0.49$), followed closely by the relations between intelligence and language courses (e.g., English, German, reading, and literature; $\rho = 0.44$), social science courses (e.g., social studies, history, and geography; $\rho = 0.43$), and art and music courses ($\rho = 0.31$). The subgroup sports ($\rho = 0.09$) differed significantly from the other domains.

In addition to such measurements of what are often high-stake academic achievements, intelligence also explains variations in learning outcomes in universities and (high) schools. This applies to predictions of long-term learning in laboratory settings, so long-term learning, as targeted in the present work, includes individuals' final test performance on learned materials as well as retrieval, retention, and memory outcomes. Hence, taking into account the importance of successful lifelong learning, a focus on intelligence for such long-term learning is inevitable. Additionally, cognitively stimulating learning environments increase the acquisition of usable knowledge, especially for learners with higher intelligence (e.g., Stern, 2015). Thus, it is also discussed that even learning situations developed to generally enhance and optimize durable and long-term learning and test performance like *desirable difficulties* (Bjork, 1994) are moderated by intelligence (e.g., Brewer & Unsworth, 2012; Kaiser, Mayer, & Malai, 2018; Minear, Coane, Boland, Cooney, & Albat, 2018).

1.1. Intelligence and desirable difficulties

It is often proposed that learning situations, methods, and processes that are difficult, challenging, and non-fluent–thus, counting as stimulating learning environments–lead to higher long-term learning, better performance, deeper information processing, and more exact recall (e.g., Bjork, 1994; Bjork & Bjork, 2011; Karpicke, Butler, & Roediger III, 2009). Recent literature describes different incantations of such desirable difficulties, two of the most common being part of the present work, respectively in Studies 1 and 2: The *generation effect* (e.g., active generation of answers to questions, generation/finding of solutions to problems, or production of information; e.g., Bertsch, Pesta, Wiscott, & McDaniel, 2007; Slamecka & Graf, 1978) and the *testing effect* (also labeled *testing, retrieval practice, test-enhanced learning*; taking learning tests or quizzes and active retrieval of information; e.g., Adesope, Trevisan, & Sundararajan, 2017; Dobson & Linderholm, 2015; Rowland, 2014; for further desirable difficulties–like *disfluency* or *distributed learning*–see: e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Weissgerber & Reinhard, 2017). Both effects increase long-term learning, test performance, and retention, and are supposed to be intricately linked, not easy to distinguish, based on and triggering the same underlying processes (e.g., active retrieval of information), and sharing a common theoretical basis.

Theoretically, the beneficial effects of testing and generation are attributed to the stimulation of cognitive processes that increase the understanding, deeper semantic/cognitive processing, and encoding of information (e.g., Bjork, 1994; Bjork & Bjork, 1992, 2011; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Rowland, 2014). Moreover, generation and testing are assumed to lead to more analytic and elaborative reasoning/thinking, more retrieval practice, more memory consolidation, more elaboration, and the allocation of more resources regarding cognition, effort, and time (e.g., Bjork, 1994; Bjork & Bjork, 1992, 2011; Dunlosky et al., 2013; Rowland, 2014). They are further supposed to strengthen memory paths, traces, associations, and the relation between stimulus and response, to help anchor the learned information in long-term memory, and to connect the retrieved information with information already stored in memory (see also *lexical activation*; e.g., Bjork & Bjork, 1992, 2011; Carpenter, 2009; Fiorella & Mayer, 2016; Gardiner & Hampton, 1985; Hirshman & Bjork, 1988). Moreover, the positive effects of desirable difficulties are higher with increased effort, quality, intensity, depth, and difficulty of retrieval (e.g., Alter, Oppenheimer, Epley, & Eyre, 2007; Bjork & Bjork, 1992; Pyc & Rawson, 2009; Rowland, 2014; Tyler, Hertel, McCallum, & Ellis, 1979).

Researchers often highlight the importance of different moderators, fulfilled requirements, or special learner characteristics, like cognitive abilities, for the effectiveness of such desirable difficulties (see also the *aptitude-treatment-interaction* and the *expertise-reversal effect*; e.g., Kalyuga, Ayres, Chandler, & Sweller, 2003; Lehmann, Goussios, & Seufert, 2016; McDaniel & Butler, 2011; Snow, 1989; Van Gog & Sweller, 2015). They also argue that testing and generation are only able to increase long-term learning and performance if the needed extended thought, further/higher effort, and the more elaborated, analytic, or effortful processing are even possible (e.g., Alter, Oppenheimer, & Epley, 2013; Oppenheimer & Alter, 2014). Higher intelligence and higher cognitive resources should increase this possibility. Further, even by definition, intelligence consists of (effective) information processing and the ability to learn, reason, and solve (complex) problems; it has also been found to be correlated with information processing, working memory capacity, and retrieval from long-term memory (e.g., Bornstein et al., 2013; Gottfredson, 1997; Oberauer, Schulze, Wilhelm, & Süß, 2005; Stern, 2015, 2017; Sternberg, 1997; Wang, Ren, & Schweizer, 2017). Intelligence is thus crucial for the theoretical basics of desirable difficulties like generation and testing.

Previous studies have empirically shown that difficult learning situations only increase long-term learning for individuals who possessed special skills (e.g., higher reading skills), sufficient cognitive resources (e.g., higher working memory capacities), or further knowledge (e.g., background/previous knowledge, experience, expertise), or for those who were in general high achieving (e.g., Carpenter et al., 2016; Kalyuga, Chandler, Tuovinen, & Sweller, 2001; Lehmann et al., 2016; McDaniel, Hines, & Guynn, 2002; McNamara, Kintsch, Songer, & Kintsch, 1996). Notably, all of these variables are linked to higher intelligence.

In line with this, the beneficial effects of desirable difficulties like generation and testing were sometimes only found for materials and learning situations that were not overly complex, not too high in element interactivity, or not too high in cognitive working memory load, as well as for information that was successfully retrieved or generated (e.g., Clark & Linn, 2003; Kaiser et al., 2018; Richland, Bjork, Finley, & Linn, 2005; Roelle & Berthold, 2017; Rowland, 2014; Van Gog & Sweller, 2015). We argue instead that these restrictions do not apply to individuals with higher intelligence. For instance, the positive and predictive effects of intelligence have been found to be especially strong regarding cognitive complex tasks and found to be especially valuable in situations that deal with more complex topics that only individuals with higher, or at least appropriate, cognitive abilities can master (e.g., Gottfredson, 1997; Kuncel et al., 2004; Roth et al., 2015). Further, when learners receive no feedback, the efficiency of testing and generation depends on the success while working on the difficult tasks: More correctly retrieved answers, which should increase with higher intelligence, lead to more long-term learning and better performance (e.g., Richland et al., 2005; Rowland, 2014).

Due to all this, it is often assumed that only individuals with sufficient cognitive abilities or resources are even able to use the

stimulations through difficult tasks to intensify and deepen their learning, in particular regarding relatively complex and realistic materials (e.g., Kalyuga et al., 2001; Lehmann et al., 2016; McDaniel et al., 2002). Only individuals with sufficient cognitive abilities are supposed to be able to correctly retrieve, further process, encode, and understand the information, even after working on difficulties that reduce processing capacities, and to ultimately manage such difficulties without being cognitively overwhelmed (e.g., Kalyuga et al., 2001; Lehmann et al., 2016; McDaniel et al., 2002). Since completing complex learning tasks, overcoming challenging difficulties, and correctly retrieving solutions or answers to questions is more likely the higher the intelligence of a learner is, we argue that intelligence moderates the effectiveness of testing and generation.

Notably, not much research has been conducted to test this hypothesis. Still, two studies that did focus on possible moderating effects of intelligence on the effectiveness of desirable difficulties found supporting evidence. For instance, Kaiser et al. (2018) tested the link between intelligence–cognitive abilities in the form of inductive figural reasoning–and generation tasks as a part of inquiry-based learning on long-term learning and successful retention of learned materials. They found that cognitive abilities increased learners' previous knowledge, which was in turn linked to higher long-term learning in the future. Minear et al. (2018) further showed that higher fluid intelligence increased the positive effect of testing for difficult, as opposed to easy, information (regarding Swahili-English word-pairs). Learners with lower fluid intelligence showed the reverse effect, thus, a stronger beneficial testing effect for easy compared to difficult information.

Nonetheless, two studies also resulted in contrary findings: Brewer and Unsworth (2012), for instance, showed testing compared to a reading control condition to be more beneficial for participants with lower general-fluid intelligence than for participants with higher general-fluid intelligence. Notably, participants with higher general-fluid intelligence were better in both learning conditions, generally performing better and having more long-term learning than participants with lower intelligence. Robey (2017) found no interaction of intelligence and a learning condition consisting of a testing and a control condition. The authors (Brewer & Unsworth, 2012; Robey, 2017) explained their results insofar as they assumed that learners with higher intelligence already used more elaboration, retrieval strategies, and cognitive processing while learning. These learners directly increased their long-term learning and performance in both conditions, in turn reducing the beneficial effects of difficult tests that arise through increased elaboration, retrieval, and cognitive processing. In contrast, the authors suppose that learners with lower abilities do not automatically use elaborative strategies and elaborative processing, so that tests, through increasing elaboration, still lead to more long-term learning. The authors further refer to the importance of increased effort while working on difficult tasks, because difficulty leads to more effort that in turn increases later learning and performance. They assumed that the learning tests were less difficult for learners with higher intelligence than for learners with lower intelligence, especially considering that word-pairs were used. This was supposed to decrease the effort more intelligent learners exerted while retrieving information, further reducing positive effects of testing. This is in line with the findings of Minear et al. (2018); their study showed that more intelligent learners had a stronger testing effect for difficult information, which probably triggered more effort and more challenges that intelligent learners still managed to overcome. In contrast, less intelligent learners had stronger testing benefits for easy information as compared to difficult information, possibly because for them easy information already triggered enough additional effort that they were still able to overcome without being overwhelmed (Minear et al., 2018).

Hence, using difficult, complex, and realistic learning materials, like those used in school or university settings, should nonetheless lead to more benefitting effects of generation and tests for learners with higher intelligence as opposed to learners with lower intelligence.

1.2. The present research

The purpose of the present work was to evaluate the effect of difficulties–generation and testing–on individuals' long-term learning. We hypothesized that the desirable difficulties conditions would yield better long-term learning than would a control condition. This main effect was, however, expected to be moderated by learners` general intelligence, insofar as that the benefits of the difficulties should be more pronounced for those with higher intelligence. We wanted to test in particular whether higher intelligence serves as a prerequisite for the beneficial effects of desirable difficulties, especially using complex, naturalistic, and curricular learning materials.

As mentioned, higher intelligence is theoretically assumed to be one precondition that elicits beneficial effects of desirable difficulties. Only learners with higher intelligence are expected to be at all able to successfully solve generation tasks, to successfully answer learning test questions, and to overcome the posed challenges. Further, individuals with higher intelligence are supposed to profit from difficult tasks in particular because they should still have cognitive resources left after solving difficulties. Because (higher) intelligence is particularly important in complex and stimulating situations, we assume that most of the restricting factors of desirable difficulties mentioned above–e.g., regarding too complex or overwhelming materials–do not apply for learners with higher intelligence. Thus, learners with higher intelligence are supposed to be able to benefit from difficulties even when learners with lower intelligence do not. We focus on intelligence as one of the most important individual characteristics concerning learning and on its supporting or essential role for difficult learning situations.

This research is important because it helps us further understand the role of cognitive processes and intelligence involved in learning in general and in the effectiveness of stimulating learning situations like desirable difficulties, and it may reveal which cognitive processes are important for difficulties like generation and testing to be desirable. It also allows us to observe if such difficult learning situations are at all able to incrementally predict long-term learning beyond intelligence since the incremental predictivity of learning methods is often supposed to be rather low (e.g., Spinath, Spinath, Harlaar, & Plomin, 2006).

Following the theoretical arguments stating intelligence as one of the best predictors for academic achievement and learning, we assumed that intelligence is positively correlated with later long-term learning (*Hypothesis 1*).

In line with the theoretical reasoning of Bjork (1994), we additionally expected a main effect of the learning condition: We assumed that participants learning with desirable difficulties like generation achieve higher long-term learning than participants in a reading control condition (*Hypothesis 2*).

We further supposed an interaction between participants' intelligence and the learning condition. We predicted the beneficial effect of the generation condition–as opposed to the reading control condition–on long-term learning to be stronger for more intellingent participants and weaker for less intelligent participants (*Hypothesis 3*).

These hypotheses will be tested in our first study. We used generation tasks (active generation/retrieval of solutions to mathematical questions without feedback) as the desirable difficulty.

2. Study 1

2.1. Methods

2.1.1. Participants

Power was set to 0.80 and sample size was calculated to detect a medium effect ($f = 0.25$). Using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009), a power analysis revealed a required sample size of $N = 128$ to detect a significant effect (alpha level of 0.05) given there is a true effect. To test the aforementioned hypotheses, we recruited a sample consisting of 150 participants ($M_{age} = 25.18$, $SD_{age} = 5.94$,

range: 19–70, 60.7% female, 86.7% German native speakers). Ninety-four percent were students at a German university studying a range of disciplines including psychology, social science, economic sciences, education, languages, history, and philosophy. One participant was excluded because he did not participate in all three sessions of the study, so the final sample consisted of $N = 149$ participants. Each participant was randomly assigned to one of the two between-subjects learning conditions: the generation condition ($n = 75$) or the reading control condition ($n = 74$). Before starting the experiment, each had to provide her or his approval through reading and signing a written informed consent.

2.1.2. Session 1

In the first session (95 min) intelligence was assessed. Up to seven participants simultaneously took part. For more anonymity and less diversion, each participant sat in a workplace with dividers.

Intelligence was measured using the German paper and pencil version of the basis module of the *Intelligence Structure Test* (I-S-T 2000 R, Liepmann, Beauducel, Brocke, & Amthauer, 2007), which includes verbal as well as non-verbal tasks. The test is based on multiple paradigms/theories regarding intelligence (e.g., Cattell, 1963, 1987; Cattell & Horn, 1978; Guttman, 1965; Horn & Cattell, 1966; Jäger, 1982; Spearman, 1939; Thurstone, 1938, 1947; see: Liepmann et al., 2007) to combine their relative advantages. The test battery of the basis module assesses a general intelligence factor, which is supposed to measure reasoning as a higher order of intelligence, and which can be interpreted as a measure of g or as fluid intelligence including knowledge. This *overall score* ($\alpha = 0.96$; further: overall IQ) consists of three areas of intellectual ability/three content factors, each assessed through three different exercises: *verbal intelligence* (*Sentence Completion*, *Verbal Analogies*, *Similarities*; $\alpha = 0.88$; further: verbal IQ), *numerical intelligence* (*Numerical Calculations*, *Number Series*, *Numerical Signs*; $\alpha = 0.95$; further numerical IQ), and *figural intelligence* (*Figure Selection*, *Cubes*, *Matrices*; $\alpha = 0.87$; further: figural IQ). Because our learning materials consist of mathematical tasks (see Session 2 below for further information), numerical IQ could be more strongly linked to long-term learning than could the other intelligence quotients; however, we plan to analyze our data using only overall IQ, due to the theoretical importance of a general mental ability and reasoning, and we plan to only use the content factors if these are more strongly correlated to participants' long-term learning than overall IQ. The intelligence scores were conducted using a standard table that took participants' age and school education into account.

Because we assessed intelligence before presenting participants the learning situation in Session 2 and before measuring their long-term learning in Session 3, we assumed a causal effect of intelligence.

2.1.3. Session 2

In the second session (60 min; at least 1 day after Session 1), demographic measures were assessed (e.g., age, gender, occupational status, native language, ethnicity, and field of study). Different control variables were also measured in randomized order to test for possible differences between participants in the generation condition and participants in the reading control condition. However, because these are not the focus of our study, we plan to only report analyses including these control variables if they differ significantly between participants in the generation condition and participants in the reading control condition.

The different control variables included participants' *Need for Cognition* (NFC; German version, short form: Bless, Wänke, Bohner, Fellhauer, & Schwarz, 1994) and participants' self-reported learning strategies and goal orientations with the German version of the *Learning and achievement motivation assessment scales* (SELLMO; Spinath, Stiensmeier-Pelster, Schöne, & Dickhäuser, 2002). We also measured participants' mathematic self-concept (German version: Schwanzer, 2002; based on: Marsh & O'Neill, 1984), their problem-solving self-concept (German version: Schwanzer, 2002; based on: Marsh & O'Neill, 1984), and their self-concept of intellectual abilities (German version: Schwanzer, 2002). Finally, we also assessed their general academic self-concept (Dickhäuser, Schöne, Spinath, & Stiensmeier-Pelster, 2002).

After that, the learning phase started; prior to that, participants were informed that they had to take a final test, about 2 weeks later, regarding the learned information. Notably, the learning situations can be classified as low-stake because participants' outcomes and performances had no influence on their actual university courses or on their everyday lives. With regards to the learning materials, we chose complex, realistic, and curricular mathematical tasks concerning regression analyses. Former research on the one hand found effects of generation tasks to be especially strong for mathematical tasks (e.g., Bertsch et al., 2007; Wirebring et al., 2015) and showed on the other hand that the linkage between intelligence and grades was especially strong for mathematical courses (Roth et al., 2015). The chosen tasks consisted of arithmetic problems (e.g., identifying a line based on two given points in a graph) and mathematical word problems (e.g., sketching of a function into a graph). First, participants' previous knowledge of such tasks was measured (duration: 4 min; one point per correct answer, at most 13 points; for a short example see Appendix A). Three independent raters evaluated and rated participants' answers to the questions in the previous knowledge test; later analyses will use the mean score of the three ratings. The inter-rater reliability was high (ICC: 0.992). Prior knowledge/expertise is often seen as a good predictor of learning outcome/long-term learning (e.g., Stern, 2015), especially when combined with intelligence, although it was not a substitute for intelligence (e.g., Grabner, Stern, & Neubauer, 2006; Stern, 2015). Moreover, previous knowledge has been shown to be important for the effectiveness of desirable difficulties (e.g., McNamara et al., 1996). Kaiser et al. (2018) for instance assumed that only learners with higher previous knowledge are able to successfully generate information and to handle the increased cognitive load. Thus, our analyses will include previous knowledge. Ultimately, we want to test if the effects of intelligence remain robust when controlling for previous knowledge.

Afterwards, in the first learning phase, all participants read basic information on the topic and saw some short examples. For the subsequent second learning phase (20 min), participants were then randomly assigned to either a generation condition or a reading control condition.

2.1.3.1. Generation condition. In the generation condition, participants had to actively fill blanks and solve mathematical wording problems, independently and actively generating and retrieving solutions to mathematical questions. There were 30 blanks overall to fill in, as well as eight solutions to generate (for a short example see Appendix A). Participants received no feedback about their generated solutions.

2.1.3.2. Reading control condition. In the reading control condition participants were presented with the same mathematical questions; however, these were already solved, and they could see the solutions (for a short example see Appendix A). Participants were instructed to read, understand, and memorize the solutions of the learning tasks.

Finally, participants reported their perception of the learning tasks, e.g., regarding the perceived difficulty of the second learning phase.

2.1.4. Session 3

Two weeks later (range: 11–19 days; $M = 14.07$; $SD = 0.75$), long-term learning (the number of correctly retrieved answers in the final test inquiring about the learned information) was assessed in a third session (70 min). Participants were required to work on final test tasks for 20 min and could gain up to 43 points (one point per correct answer). Participants' performance, indicating their long-term learning, was also coded from the same three independent raters. Later analyses will use the mean score of the three ratings. Inter-rater reliability for long-term learning was high as well (ICC: 0.897). The final test tasks again included arithmetical problems (35 blanks to fill/to answer), as well as mathematical text or word problems (eight problems to solve) and were similar to the learning tasks in the previous session. Some

questions, however, also included equations with negative slopes in contrast to the completely positive slopes presented in the learning phase. Moreover, the final test was also a low-stake testing situation because participants knew that there were no consequences with respect to their performance; neither their rewards for participating nor their regular university courses and grades were dependent of their final test performance and long-term learning outcome.

We further assessed participants' perception of the final test tasks and if they had gathered further information on the learned topic in the intervening time. Following an unrelated study concerning credibility judgments, participants then answered final questions about our study, e.g., regarding thoughts and comments. They also had the opportunity to subscribe to a post-experimental elucidation, were shortly debriefed, received 25 Euro as a reward (psychology students earned course credits instead), and could review their intelligence quotients.

2.2. Results

On average, participants achieved an overall IQ of 98.54 (SD = 14.58, range: 73.0–131.5). Participants had on average a verbal IQ of 100.48 (SD = 14.51, range: 70.0–130.0), a numerical IQ of 95.79 (SD = 16.29, range: 56.5–133.0), and a mean figural IQ of 97.72 (SD = 14.83, range: 61.0–137.5). In the previous knowledge test, participants were on average able to correctly answer 7.79 of the 13 (59.9%) questions (SD = 4.17, range: 0–13).

Gender distribution, intelligence, previous knowledge, the time lag between Session 2 and Session 3, and the aforementioned control variables did not differ between participants in the generation condition and participants in the reading control condition (all $ps \geq .438$). Therefore, our analyses will not include the control variables. Regarding the manipulation check, there was no significant difference between participants' ratings of the difficulty of the learning tasks between the two learning conditions ($M_{generation} = 2.56$, $SD_{generation} = 1.08$, $M_{reading} = 2.65$, $SD_{reading} = 1.07$, $t(147) = -0.50$, $p = .615$, $d = -0.08$; range: 1–5). For the following analyses, we z-standardized the predictors and used Process (Hayes, 2018). In the regression analyses we further report the semi-partial correlations ($r_{y(x,z)}$).

2.2.1. Intelligence, learning condition, and long-term learning

Considering the final test tasks measuring long-term learning, participants were on average able to correctly answer 22.04 of 43 (51.3%) final test questions (SD = 10.73, range: 0.67–42.33).

Correlations (not corrected as well as corrected for measurement error, thus, *disattenuated* correlations) can be seen in Table 1. As expected, overall IQ was significantly correlated to participants' later long-term learning ($r = 0.67$, $p < .001$, disattenuated correlation: $r = 0.72$). In line with the expected validity of the scale, overall IQ was also correlated significantly to the three content factors (see Table 1). Interestingly, overall IQ and numerical IQ were strongly correlated ($r = 0.84$, $p < .001$, disattenuated correlation: $r = 0.88$), indicating these variables to be extremely linked and almost identical. Furthermore, the positive correlation between numerical IQ and long-term learning ($r = 0.66$, $p < .001$, disattenuated correlation: $r = 0.72$) did not significantly differ from the correlation of overall IQ and long-term learning ($r = 0.67$, $p < .001$, disattenuated correlation: $r = 0.72$; $z = 0.30$, $p = .384$). Thus, we use overall IQ as the predictor for the following analyses. Previous knowledge also positively correlated to overall IQ as well as to participants' later long-term learning (see Table 1).

Regarding Hypothesis 1, the significant correlation between overall IQ and long-term learning ($r = 0.67$, $p < .001$, disattenuated correlation: $r = 0.72$; see Table 1) found first support for our assumption. To test Hypothesis 2, we conducted a t-test to compare the average long-term learning of participants in both learning conditions: $M_{generation} = 21.13$, $SD_{generation} = 10.66$, $M_{reading} = 22.97$, $SD_{reading} = 10.79$, $t(147) = -1.05$, $p = .297$, $d = -0.17$. Contrary to Hypothesis 2, there was no significant difference in long-term learning between participants in the reading control condition and participants in the generation condition.

Table 1
Correlations among intelligence, previous knowledge, and long-term learning (N = 149).

	1.	2.	3.	4.	5.	6.
1. Overall IQ		0.69	0.88	0.84	0.53	0.72
2. Verbal IQ	0.63**		0.36	0.42	0.36	0.45
3. Numerical IQ	0.84**	0.33**		0.52	0.46	0.72
4. Figural IQ	0.77**	0.37**	0.47**		0.40	0.44
5. Previous knowledge	0.52**	0.34**	0.45**	0.37**		0.67
6. Long-term learning	0.67**	0.40**	0.66**	0.39**	0.63**	

Note: The (uncorrected) correlations are displayed below the diagonal, the disattenuated correlations are presented above the diagonal.
** $p < .001$.

To test Hypothesis 1 and Hypothesis 2 in a more detailed way, we conducted a linear regression analysis using the learning condition (0 = reading control condition, 1 = generation condition) and participants' overall IQ as predictors for long-term learning. Homoscedasticity was given (Breusch-Pagan-Test: $p = .477$). R for this regression was significantly different from zero, $F(2,146) = 59.85$, $R^2 = 0.451$, $R^2_{adj} = 0.443$, $p < .001$. Again, contrary to our expectation, the learning condition was not a significant predictor of long-term learning, $t(146) = -1.52$, $B = -2.00$, $SE = 1.31$, $\beta = -0.09$, $p = .130$, $r_{y(x,z)} = -0.093$. Overall IQ did, as expected, show a significant effect in the equation, $t(146) = 10.85$, $B = 7.14$, $SE = 0.66$, $\beta = 0.67$, $p < .001$, $r_{y(x,z)} = 0.666$. We further ran another regression analysis to control for previous knowledge. Homoscedasticity was given (Breusch-Pagan-Test: $p = .177$). R for this regression was significantly different from zero, $F(3,145) = 60.83$, $R^2 = 0.557$, $R^2_{adj} = 0.547$, $p < .001$. The regression also significantly explained more variance than the model without previous knowledge, $F_{change} = 34.95$, $p < .001$. Learning condition was still not a significant predictor of later long-term learning in the final test, $t(145) = -1.14$, $B = -1.35$, $SE = 1.19$, $\beta = -0.06$, $p = .256$, $r_{y(x,z)} = -0.063$. As expected, overall IQ was still a significant predictor, $t(145) = 7.14$, $B = 4.98$, $SE = 0.70$, $\beta = -0.46$, $p < .001$, $r_{y(x,z)} = 0.395$. Previous knowledge was also a significant predictor of long-term learning, $t(145) = 5.91$, $B = 4.13$, $SE = 0.70$, $\beta = 0.39$, $p < .001$, $r_{y(x,z)} = 0.327$. Thus, controlling for previous knowledge did not substantially change the effects. Hence, Hypothesis 1 can be supported but Hypothesis 2 cannot.

To test Hypothesis 3, we added the interaction-term of overall IQ and the learning condition as a predictor. Homoscedasticity was given (Breusch-Pagan-Test: $p = .172$). R for this regression was significantly different from zero, $F(4,144) = 45.62$, $R^2 = 0.559$, $R^2_{adj} = 0.547$, $p < .001$. The regression did not explain more variance than the first model, $F_{change} = 0.56$, $p = .457$. The learning condition was again not a significant predictor, $t(144) = -1.15$, $B = -1.37$, $SE = 1.19$, $\beta = -0.06$, $p = .253$, $r_{y(x,z)} = -0.064$. As expected, overall IQ was significant, $t(144) = 4.56$, $B = 4.46$, $SE = 0.98$, $\beta = 0.42$, $p < .001$, $r_{y(x,z)} = 0.252$. Previous knowledge was also a significant predictor, $t(144) = 5.83$, $B = 4.09$, $SE = 0.70$, $\beta = 0.38$, $p < .001$, $r_{y(x,z)} = 0.323$. The interaction of both variables showed no significant effect in the equation, $t(144) = 0.75$, $B = 0.90$, $SE = 1.21$, $\beta = 0.07$, $p = .457$, $r_{y(x,z)} = 0.041$. Thus Hypothesis 3 was not supported.

2.3. Discussion

Our results supported Hypothesis 1 but not Hypotheses 2 and 3. As expected, intelligence was significantly and positively linked to later long-term learning indicated by participants' performance in the final test assessing the learned information. This fits the aforementioned theoretical reasoning that claims intelligence as one of the strongest predictors for long-term learning and academic achievement (e.g.,

Fergusson et al., 2005; Kuncel et al., 2004; Roth et al., 2015; Spinath et al., 2006). Thus, our findings again strengthen the importance of intelligence for success in the academic field. Notably, the positive effect of intelligence for later long-term learning remained robust when adding participants' previous knowledge, thus, the benefitting effects of intelligence were not due to specific knowledge. Still, adding previous knowledge increased the explained variance.

Contrary to theoretical argumentations and former research (e.g., Bertsch et al., 2007; Bjork, 1994; Slamecka & Graf, 1978) generation in comparison to reading did not lead to more long-term learning. Taking a closer look at the manipulation of the learning condition, however, showed no significant differences between participants' ratings of the difficulty of the two learning conditions. Participants in the generation condition did not perceive the generation tasks as more difficult than participants in the reading control condition perceived the reading tasks. We argue, then, that the generation tasks were not difficult enough to count as a desirable difficulty and were in turn unable to trigger the needed increase in effort, elaboration, retrieval, and cognitive processing. Due to our questioning of the successfulness of the manipulation, learning condition differences must be more pronounced in future studies. Thus, more difficult learning conditions should actually be more difficult.

We furthermore found no interaction between intelligence and the learning condition. This is plausible considering that there were no main effects of the learning condition.

Our sample had an extremely large range of intelligence scores and a–at first glance–rather low mean overall intelligence score of 98.54, which seemed surprising for a university sample. However, none of the participants was an outlier, and the calculation of the intelligence scores took into direct account participants' educational background, so ultimately the intelligence scores proved normed for individuals with the educational level required to study in Germany. In line with this, a large number of individuals per generation these days decides to attend university: In one survey, for instance, 74% of those who were entitled to study due to their school degree had already begun studying within 6 months post-graduation, or they definitely planned to begin soon (Schneider, Franke, Woisch, & Spangenberg, 2017). Only a small number of individuals with this educational background wanted to instead start an apprenticeship or vocational education (about 20%; Schneider et al., 2017). Moreover, many study paths in Germany do not pose entry restrictions or pre-requisites, so studying individuals in general and in our sample were not highly selected by admission procedures. Due to that, samples of students are more diverse regarding their cognitive resources; are argued to have wider ability ranges; are generally not restricted to higher abilities; and should resemble the general population. Further, the average intelligence score of our sample did not differ from the expected population average of 100 ($t(148) = -1.222$, $p = .224$, $d = -0.10$). Moreover, almost all participants stated that they were interested in the results of their intelligence test and later reviewed these. Hence, we think that our sample adequately represents students in the observed university.

3. Study 2

The aim of Study 2 was to again test the three hypotheses stated above. That said, the method of Study 2 is parallel to the method used in Study 1, except we implemented learning tests with feedback as the desirable difficulty as opposed to a re-reading control condition. The study also had two instead of three sessions. We further used different learning materials and measured long-term learning 1 week instead of 2 weeks after learning.

We again assumed that intelligence is positively linked to participants' later long-term learning (*Hypothesis 1*).

We additionally supposed that participants in the desirable difficulties learning condition using tests achieve higher long-term learning than do participants in the re-reading control condition (*Hypothesis 2*).

We further assumed that the beneficial effect of the testing condition is moderated by intelligence, insofar as that the positive effect is stronger for more intelligent and weaker for less intelligent participants (*Hypothesis 3*).

3.1. Methods

3.1.1. Participants

Power was set to 0.90 and sample size was calculated to detect a medium effect ($f = 0.25$). Using G*Power (Faul et al., 2009), a power analysis revealed a required sample size of $N = 171$ to detect a significant effect (alpha level of 0.05) given there is a true effect. To test the aforementioned hypotheses, we recruited a sample consisting of 179 participants. Three participants were excluded because they did not participate in both sessions of the study, so the final sample consisted of $N = 176$ participants ($M_{age} = 22.83$, $SD_{age} = 4.23$, range:18–40, 70.5% female, 84.1% German native speakers). Ninety-six-point-six percent were students at a German university studying a range of disciplines including psychology, social science, economic sciences, education, physics, mathematics, languages, history, and philosophy. Each participant was randomly assigned to one of the two between-subjects learning conditions: the testing condition ($n = 87$) or the re-reading control condition ($n = 89$). Before starting the experiment, each had to provide her or his approval through reading and signing a written informed consent.

3.1.2. Session 1

In the first session (135 min) intelligence was assessed and the learning phase took place. The laboratory and the setting were the same as in Study 1.

We measured intelligence using the same test as before (I-S-T 2000 R, Liepmann et al., 2007). Again, due to the theoretical importance of a general mental ability and the results of Study 1, we analyze our data using overall IQ. After a short break, we then assessed demographic measures of the participants (e.g., age, gender, occupational status, native language, ethnicity, and field of study).

Before the learning phase started, participants were informed that they would, 1 week later, be charged with taking a final test covering the learned information. The learning material consisted of a textbook chapter on the brain's lateralization based on a standard introductory textbook often adopted for university courses in biopsychology (Pinel & Pauli, 2012). This material was complex, curricular, and difficult; moreover, previous research found the linkage between intelligence and grades to be especially high for biological courses (e.g., Roth et al., 2015).

Again, due to the above-mentioned theoretical importance of previous knowledge and the results of Study 1, participants' previous knowledge was measured by implementing three open-ended questions (duration: 3 min; one point per correct answer, at most three points; for a short example see Appendix B). Three independent raters evaluated and rated participants' answers to the previous knowledge questions: Analyses will use the mean score of the three ratings. The inter-rater reliability was high (ICC: 0.983).

In the first learning phase, all participants had 10 min to once read three pages of the textbook concerning the brain's lateralization. For the subsequent 10 min of the second learning phase, participants were then randomly assigned to either a testing condition or a re-reading control condition.

3.1.2.1. Testing condition. In the testing condition, participants were presented with a learning test assessing aspects of the previously read textbook material. The 17 test questions consisted of multiple-choice questions, each with one correct answer and three distractors, plus open-ended questions asking for single words or bullet points but also requiring longer, more detailed answers (a maximum of 20 points could be gained; for examples see Appendix B). After spending 9 min

answering the questions of the learning test, participants received feedback in the form of an answer sheet displaying the correct answers of the learning test (see Rowland, 2014, concerning the importance of feedback for the effectiveness of tests).

3.1.2.2. Re-reading control condition. In the re-reading control condition participants were again presented with the same textbook materials. They were instructed to read the text as many times as they wanted in the given time and to learn, understand, and memorize the information as best as they could.

Finally, participants answered some manipulation check questions regarding this second part of the learning phase, e.g., regarding the difficultly, strenuousness, or effectivity of the learning task, or regarding participants' perceived success.

3.1.3. Session 2

In the second session (about 70 min; 1 week after Session 1; range: 7–9 days; $M = 7.02$, $SD = 0.18$) participants' long-term learning was measured. Participants were therefore required to work for 10 min on a final test that included 22 final test questions. They could gain up to 27 points (between one and two points could be achieved per correctly answered final test question). In line with the learning test in the testing condition, the final test consisted of multiple choice and open-ended questions. Some final test questions were identical to questions used in the learning test, some questions were changed slightly, and some were part of the previously read textbook but were not previously implemented in the learning test. The three independent raters again evaluated and rated all answers on the final test (ICC: 0.973); analyses will use the mean score of these three ratings. Again, the final test was a low-stake situation because there were no consequences for participants dependent on their performance.

We further assessed participants' perception of the final test, e.g., regarding the difficulty or strenuousness of the final test and if they had gathered further information on the learned topic in the intervening time. Following an unrelated study concerning credibility judgments, participants answered final inquiries about our study, e.g., regarding thoughts and comments. They also had the opportunity to subscribe for a post-experimental elucidation, were shortly debriefed, received their 25 Euro reward (psychology students earned course credits instead), and could review their intelligence quotients.

3.2. Results

On average, participants achieved an overall IQ of 101.03 ($SD = 14.64$, range: 67.5–137.5). They had on average a verbal IQ of 99.76 ($SD = 13.34$, range: 69.0–133.0), a numerical IQ of 98.71 ($SD = 15.98$, range: 66.0–133.0), and a figural IQ of 100.48 ($SD = 14.75$, range: 79.5–140.5). Participants achieved on average 0.27 of 3 points in the previous knowledge test ($SD = 0.61$, range: 0–3).

Participants' age, gender distribution, native language distribution, the time lag between Session 1 and Session 2, intelligence, and previous knowledge did not significantly differ between participants in the testing condition and participants in the re-reading control condition (all $ps \geq .317$). Regarding the manipulation check, participants in the testing condition rated the learning task as significantly more difficult than participants in the re-reading control condition ($M_{testing} = 3.09$, $SD_{testing} = 0.71$, $M_{re\text{-}reading} = 2.33$, $SD_{re\text{-}reading} = 0.78$, $t(174) = 6.81$, $p < .001$, $d = 1.02$). More participants in the testing condition stated that they would need further time to work on the learning task (32 vs. 3 participants). In line with this, participants in the re-reading control condition perceived themselves as significantly more successful while working on the learning task than did participants in the testing condition ($M_{testing} = 2.85$, $SD_{testing} = 0.93$, $M_{rereading} = 3.42$, $SD_{re\text{-}reading} = 0.77$, $t(174) = -4.93$, $p < .001$, $d = -0.67$). However, there were no significant differences between the perceived strenuousness of the testing and the re-reading condition ($M_{testing} = 3.02$, $SD_{testing} = 1.02$, $M_{re\text{-}reading} = 2.73$, $SD_{re\text{-}reading} = 1.09$, $t(174) = 1.84$, $p = .067$, $d = 0.28$) and the rated effectivity of the learning tasks ($M_{testing} = 3.72$, $SD_{testing} = 0.86$, $M_{re\text{-}reading} = 3.58$, $SD_{re\text{-}reading} = 0.95$, $t(174) = 1.02$, $p = .308$, $d = 0.16$). For the following analyses, we z-standardized the predictors and used Process (Hayes, 2018). In the regression analyses we further report the semi-partial correlations ($r_{y(x,z)}$).

3.2.1. Intelligence, learning condition, and long-term learning

Considering the final test task measuring long-term learning, participants were on average able to give 13.18 of 27 (48.8%) correct answers in the final test ($SD = 4.78$, range: 2.33–25.00). Correlations (not corrected as well as disattenuated) can be seen in Table 2. As expected, overall IQ was significantly correlated to participants' later long-term learning ($r = 0.44$, $p < .001$, disattenuated correlation: $r = 0.46$). In line with the expected validity of the scale, overall IQ was also significantly correlated to the three content factors (see Table 2). The correlation of verbal IQ and later long-term learning ($r = 0.51$, $p < .001$, disattenuated correlation: $r = 0.55$) did not significantly differ from the correlation of overall IQ and later long-term learning ($r = 0.44$, $p < .001$, disattenuated correlation: $r = 0.46$; $z = 1.25$, $p = .105$). In addition, previous knowledge was positively correlated to overall IQ as well as to participants' later long-term learning (see Table 2).

Regarding Hypothesis 1, the significant correlation between overall IQ and long-term learning ($r = 0.44$, $p < .001$, disattenuated correlation: $r = 0.46$; see Table 2) found first support for our assumption. To test Hypothesis 2, we first conducted a *t*-test comparing the long-term learning between participants in the testing condition ($M = 13.92$, $SD = 5.02$) and participants in the re-reading control condition ($M = 12.45$, $SD = 4.43$). This difference was significant ($t(174) = 2.07$, $p = .040$, $d = 0.31$), indicating a beneficial effect of testing on later long-term learning, hence supporting our second hypothesis.

To test both hypotheses in a more detailed way, we conducted a linear regression analysis with both the learning condition (0 = re-reading control condition, 1 = testing condition) and overall IQ as predictors for later long-term learning. Homoscedasticity was given (Breusch-Pagan-Test: $p = .718$). *R* for this regression was significantly different from zero, $F(2,173) = 24.92$, $R^2 = 0.224$, $R^2_{adj} = 0.215$, $p < .001$. As expected in Hypothesis 1, overall IQ showed a significant effect in the equation, $t(173) = 6.67$, $B = 2.14$, $SE = 0.32$, $\beta = 0.447$, $p < .001$, $r_{y(x,z)} = 0.447$. As expected in Hypothesis 2, the learning condition was also a significant predictor, $t(173) = 2.58$, $B = 1.65$, $SE = 0.64$, $\beta = 0.173$, $p = .011$, $r_{y(x,z)} = 0.173$. To further control for participants' previous knowledge, we ran another regression analysis including previous knowledge as a predictor. Homoscedasticity was given (Breusch-Pagan-Test: $p = .412$). *R* for this regression was significantly different from zero, $F(3,172) = 31.89$, $R^2 = 0.357$, $R^2_{adj} = 0.346$, $p < .001$. The regression significantly explained more variance than the model without previous knowledge, $F_{change} = 35.821$, $p < .001$. As expected, the learning condition was a significant predictor of later long-term learning, $t(172) = 2.57$, $B = 1.50$, $SE = 0.58$,

Table 2
Correlations among intelligence, previous knowledge, and long-term learning ($N = 176$).

	1	2	3	4	5	6
1. Overall IQ		0.71	0.88	0.86	0.21	0.46
2. Verbal IQ	0.65**		0.36	0.41	0.25	0.55
3. Numerical IQ	0.84**	0.33**		0.57	0.11	0.26
4. Figural IQ	0.79**	0.36**	0.52**		0.10	0.33
5. Previous knowledge	0.20**	0.23**	0.11	0.09		0.46
6. Long-term learning	0.44**	0.51**	0.25**	0.30**	0.45**	

Note: The (uncorrected) correlations are displayed below the diagonal; the disattenuated correlations are presented above the diagonal.

** $p \leq .001$.

β = 0.158, p = .011, $r_{y(x,z)}$ = 0.157. Overall IQ also showed a significant effect in the equation, $t(172) = 5.97$, $B = 1.78$, $SE = 0.30$, β = 0.373, $p < .001$, $r_{y(x,z)} =$. 365. Previous knowledge was also a significant predictor, $t(172) = 5.99$, $B = 1.78$, $SE = 0.30$, β = 0.374, $p < .001$, $r_{y(x,z)} = 0.366$. Thus, controlling for previous knowledge did not substantially change the effects. This again supports Hypothesis 1 and Hypothesis 2.

To test Hypothesis 3, we further added the interaction-term of overall IQ and the learning condition to this linear regression analyis. Homoscedasticity was given (Breusch-Pagan-Test: p = .338). R for this regression was significantly different from zero, $F(4,171) = 25.65$, $R^2 = 0.375$, $R^2_{adj} = 0.360$, $p < .001$. The regression significantly explained more variance than the model without the interaction, $F_{change} = 4.81$, p = .030. In line with the results above, the learning condition was a significant predictor, $t(171) = 2.60$, $B = 1.50$, $SE = 0.58$, β = 0.158, p = .010, $r_{y(x,z)} = 0.157$. Overall IQ also showed a significant effect in the equation, $t(171) = 2.89$, $B = 1.17$, $SE = 0.41$, β = 0.245, p = .004, $r_{y(x,z)} = 0.175$. Previous knowledge also was a significant predictor, $t(171) = 6.19$, $B = 1.83$, $SE = 0.30$, β = 0.383, $p < .001$, $r_{y(x,z)} = 0.374$. As expected, the interaction of overall IQ with the learning condition was also able to significantly predict later long-term learning, $t(171) = 2.19$, $B = 1.27$, $SE = 0.58$, β = 0.183, p = .030, $r_{y(x,z)} = 0.133$ (see Fig. 1). A closer look at the conditional effects of the ordinal interaction revealed that there was no significant effect of the learning condition on long-term learning for participants with rather low overall IQ (overall IQ 1SD below mean), $t(171) = 0.28$, $B = 0.23$, $SE = 0.82$, p = .780. However, participants with average overall IQ benefitted significantly from being in the testing condition compared to being in the re-reading control condition, $t(171) = 2.60$, $B = 1.50$, $SE = 0.58$, p = .010. The positive effect of the learning condition was especially strong for more intelligent participants (overall IQ 1SD above mean), $t(171) = 3.39$, $B = 2.77$, $SE = 0.82$, p = .001.[1] These findings–that intelligence moderated the effectiveness of the testing condition–supported Hypothesis 3.

Notes. N = 176. Overall IQ +1SD = overall IQ 1SD above mean; Mean overall IQ = average overall IQ; overall IQ -1SD = overall IQ 1SD below mean.

To conclude, we further added the interaction-term of previous knowledge and the learning condition to the linear regression model.[2] Again, homoscedasticity was given (Breusch-Pagan-Test: p = .338). R for this regression was significantly different from zero, $F(5,170) = 20.40$, $R^2 = 0.375$, $R^2_{adj} = 0.357$, $p < .001$, but did not explain more variance than the model without the interaction of previous knowledge and the learning condition, $F_{change} < 0.001$, p = .993. Adding this further interaction-term did not change the results: The learning condition remained a significant predictor, $t(170) = 2.59$, $B = 1.50$, $SE = 0.58$, β = 0.158, p = .010, $r_{y(x,z)} = 0.157$. Overall IQ still showed a significant effect in the equation, $t(170) = 2.83$, $B = 1.17$, $SE = 0.41$, β = 0.245, p = .005, $r_{y(x,z)} = 0.172$. Previous knowledge remained a significant predictor, $t(170) = 4.40$, $B = 1.83$, $SE = 0.42$, β = 0.383, $p < .001$, $r_{y(x,z)} = 0.267$. The interaction of overall IQ and the learning condition was also still able to significantly predict later long-term learning, $t(170) = 2.15$, $B = 1.27$, $SE = 0.59$, β = 0.183, p = .033, $r_{y(x,z)} = 0.130$. The conditional effects of this ordinal interaction thereby followed the same pattern as described before. In contrast, the interaction between the learning condition and previous knowledge was not significant, $t(170) = -0.01$, $B = -0.01$, $SE = 0.59$, β = −0.001, p = .993, $r_{y(x,z)} = -0.001$.

[1] The Johnson-Neyman region of significance for the moderator (conducted with Process) showed that the testing condition had no significant effect on long-term learning for participants with a (standardized) overall IQ below −0.256. For participants with a (standardized) overall IQ above −0.256, the testing condition had a positive and significant effect compared to a re-reading control condition.

[2] We thank an anonymous reviewer for this suggestion.

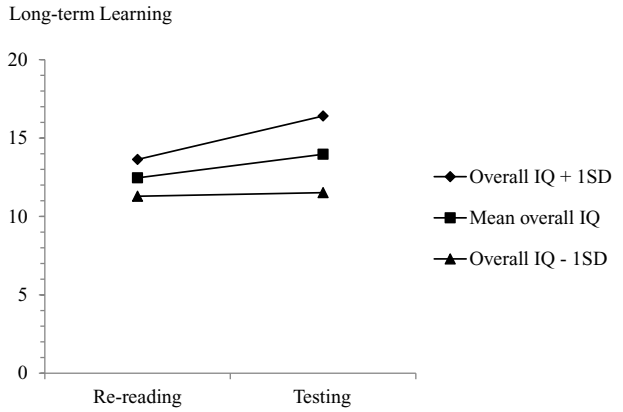


Fig. 1. Long-term learning predicted through the learning condition, participants' overall IQ, and the interaction of both variables.

3.3. Discussion

The results described above are completely in line with our theoretical predictions and thus support all three of our hypotheses. Again, as expected, higher intelligence was significantly linked to more long-term learning, highlighting the importance of (general) intelligence for long-term learning, knowledge acquisition, and academic achievement (e.g., Fergusson et al., 2005; Roth et al., 2015; Stern, 2015).

Notably, the manipulation of the learning condition–testing versus re-reading control–was successful in this second study: The testing condition was perceived as significantly more difficult than the re-reading control condition. In line with this, we now found a significant testing effect on participants' long-term learning. Participants in the testing condition retrieved more answers correctly in the final test assessing long-term learning after 1 week than did participants in the re-reading control condition. This fits theoretical assumptions regarding the beneficial effects of implementing learning tests (e.g., Adesope et al., 2017; Rowland, 2014). Interestingly, this positive effect of the learning test was found using realistic, difficult, and curricular materials that consisted of rather complex information. Thus, our results are an indication for the robustness of the testing effect, even controlling for intelligence and previous knowledge.

Next, we found a significant interaction between the learning condition and participants' overall IQ: Intelligence moderated the effectiveness of learning tests, insofar as that participants with relatively low intelligence, that is, overall IQ one standard deviation below mean, did not benefit from the manipulation of the learning situation. Thus, for lower intelligence, there was no difference in later long-term learning between participants in the testing condition and participants in the re-reading control condition. However, participants with average intelligence benefitted from taking learning tests as opposed to re-reading the same information. This positive effect was even stronger for more intelligent participants with overall IQ one standard deviation above mean; thus, more intelligent participants in particular profited from such difficult learning. This result is in line with the aforementioned theoretical assumptions and empirical findings (e.g., Alter et al., 2013; McDaniel et al., 2002; Minear et al., 2018; Oppenheimer & Alter, 2014).

Finally, participants previous knowledge was also a significant and positive predictor for long-term learning. The interaction of previous knowledge and the learning condition was, however, not significant–even though previous knowledge was found to be a moderator for the beneficial effects of desirable difficulties in past studies (e.g., McNamara et al., 1996).

4. General discussion

In two studies, we analyzed the linkage between participants' intelligence and their long-term learning, as well as moderating effects of

intelligence on difficult learning situations like generation and testing that are supposed to increase long-term learning. As mentioned in the introduction, intelligence has often been assumed to be one of the best predictors for learning and academic achievement, especially regarding complex and stimulating learning. Higher intelligence was further discussed to increase the effectiveness of intentionally hindered and more difficult learning situations and to be linked to better and more effortful cognitive information processing.

The results of our two studies highlight the importance of general intelligence and the inevitability of focusing on intelligence for predicting long-term learning. The positive linkage of intelligence and long-term learning remained robust and strong when controlling for participants' previous knowledge and when manipulating the learning situation. Moreover, although desirable difficulties, at least regarding tests in our second study, were also beneficial, intelligence even moderated the effectiveness of such difficult learning. This moderation effect regarding complex and difficult information is the most important contribution of our second study to the existing intelligence literature. Notably, tests were not more effective than re-reading control tasks for participants with relatively low intelligence but were beneficial for average and highly intelligent participants. Highly intelligent learners profited especially from using learning tests. Hence, intelligence was not only generally linked to long-term learning but also moderated situations, processes, and methods that were specifically constructed to increase long-term learning. This is in line with the above-mentioned theories stating the importance of a general intelligence factor for learning, success, and academic achievement in different contexts (e.g., Kuncel et al., 2004; Roth et al., 2015; Spearman, 1904). Additionally, our results are similar to previous (controversial) research stating educational interventions and learning methods to be especially–or even only–advantageous for individuals with at least average cognitive abilities like intelligence: Thus, methods trying to improve long-term learning and academic achievement for everyone are often suggested to only further increase the disparity between high and low ability learners (see also the *Matthew* or *rich-get-richer* effects; e.g., Rapport, Brines, Theisen, & Axelrod, 1997; Stern, 2015, 2017; Walberg & Tsai, 1983). Our results further support the literature assuming the importance of higher cognitive abilities for the beneficial effects of desirable difficulties (e.g., Kaiser et al., 2018; McDaniel et al., 2002; Minear et al., 2018). Our findings present a unique contribution to the understanding of the role of intelligence for learning in general, as well as for stimulating learning situations using difficult, challenging, and complex materials. Thus, at least average and higher intelligence facilitates effective deeper semantic encoding, cognitive processing, cognitive effort, and consolidation of information that is triggered by tests.

Due to our results, we can advise the implementation of learning tests for university students, at least for averagely and highly intelligent learners. These profit from using difficult learning tests, even when applying a rather short, low-stake test only once. Fortunately, such learning tests are advantageous for a larger population of university students and can be implemented easily into university courses. Still, lecturers must remain vigilant that the applied learning tests are actually difficult and complex enough to trigger the beneficial effects. Concerning relatively unintelligent learners, we cannot unconditionally advise lecturers to use tests because such learners would have to indulge in difficult learning without profiting from it. Nonetheless, we also cannot advise against using difficult tests because at the very least, participants with lower intelligence suffered no disadvantages on their long-term learning due to the application of learning tests (see also the often assumed *poor-get-poorer* effect; e.g., Stanovich, 1986). However, one might also argue that difficult learning is correlated with stress or frustration for less intelligent learners, because difficult tasks were in general found to increase perceived anxiety, and even low-stake quizzes were linked to pressure compared to a re-reading control task (e.g., Hinze & Rapp, 2014; O'Neil, Spielberger, & Hansen, 1969). Regarding generation tasks, implications are not that clear because the manipulation of the learning condition in Study 1 was unsuccessful. In line with this, Study 1 did not result in a significant effect of the learning condition, thus, generation was not more beneficial than a reading control task. At the very least the generation tasks did not reduce participants' long-term learning, thus, they were not harmful.

There were some positive and negative aspects of our studies that we care to mention and that could be applied or adapted in future work. For instance, the intelligence test we used was a rather detailed one with high quality factors; future research should use similar measures. This applies especially to the importance and predictivity of a general intelligence factor. Still, we only used the basis module of the intelligence test, which measures a general intelligence factor similar to g or to fluid intelligence encompassing knowledge components. Future studies may add the existing knowledge tests to additionally assess fluid and crystalline intelligence so that more information regarding intelligence is available. Both of our studies used different curricular and realistic learning materials that are actually used in school and university courses; that said, the results can be generalized for actual learning materials and for information that is complex and difficult instead of relatively abstract learning of word pairs, vocabulary, or associations. It is vital that the difficult learning tasks are perceived as more difficult than the easier control tasks and that both conditions are clearly distinguishable.

As a limitation, we only observed the influence of a single manipulated learning condition–one generation task or one learning test–on one single final test assessing long-term learning. However, it is important to test if the moderating effects of intelligence remain the same when applying multiple learning tests or multiple re-reads over the course of an entire semester. In line with this, future studies should use multiple follow-up final tests to check if the effects change over time. Although the positive effect of intelligence was found in previous studies over long periods, the beneficial effect of tests could decline. One main limitation of our studies is that in regard to intelligence, we were only able to observe correlations. Although we did infer causal effects due to the different times of measurements of intelligence and long-term learning, further causal analyses are still advantageous. Future studies should implement longitudinal designs because these are supposed to serve as a basis for causal effects (cf., Strenze, 2007, 2015).

All in all, there remain open questions regarding the tested linkage among intelligence, cognitive processes, generation, testing, and long-term learning. This applies for instance to the underlying effects of cognitive processing for learning. Although we argue that intelligence is positively correlated to better retrieval as well as to deeper processing of information, and although we know that higher intelligence is generally important for learning, we do not know exactly why. The same applies to the consideration of why desirable difficulties increase cognitive processes that lead to higher long-term learning. It is possible that higher working memory capacities, the ability to handle simultaneously more pieces of information, the amount of cognitive resources, or higher memory skills are responsible for increased long-term learning. However, higher success could also be due to the abilities to reason, abstract thinking, or elaboration, or to higher processing speed, or simply to the ability to handle more cognitive effort and to overcome challenging tasks. So, in addition to general intelligence, future studies could focus on the linkage between even more aspects of cognitive abilities, like processing speed, working memory capacity, memory, or reasoning, on long-term learning and the effectiveness of generation/testing. Moreover, future work should also focus on increasing the benefit of desirable difficulties for learners with all–and especially lower–ability levels and not only for average or highly intelligent individuals. Thus, future studies may try to design difficulties that are adequately difficult for every individual; the tasks should be difficult enough to elicit the beneficial effects of desirable difficulties but still easy enough that learners with lower intelligence are able to overcome them without being completely overwhelmed (see e.g., Minear et al., 2018). Future studies should therefore monitor and test which level of difficulty is beneficial for which individual. Lecturers could, for instance, also give lower ability learners more time or apply graded learning aids to support them (see e.g., Hänze, Schmidt-Weigand, &

Stäudel, 2010). Besides, researchers could test if lower ability learners would benefit from longer initial learning phases or from applications of desirable difficulties later in the learning process when these learners have already mastered some of the basic information or formed sufficient previous knowledge (see also the above-mentioned expertise-reversal effect or the aptitude-treatment-interaction; e.g., Kalyuga et al., 2003; Snow, 1989). Future work could also test if multiple applications of desirable difficulties or the usage of tests in high-stake learning situations in actual university courses may improve long-term learning for lower ability individuals.

In general, future work could also use a more natural setting, a within-subject design, or it could even implement further difficulty nuances regarding the information as well as the desirable difficulties themselves. Although the forced application of learning tasks is rather common in university courses, it is advantageous to explore the effects of intelligence and desirable difficulties using self-regulated learning. Thus, one could explore if intelligence also moderates the decision to use generation tasks or tests instead of relatively easy re-reading tasks, and also if intelligence moderates learners' persistence while working on such difficulties.

Conclusion

In summary, we want to emphasize the importance of intelligence: Studies 1 and 2 showed that higher intelligence was beneficial for long-term learning, even controlling for participants' previous knowledge. Study 2 also found a positive effect of difficult learning tests as opposed to the application of reading control tasks. Notably, this beneficial effect was moderated by intelligence: In particular participants with higher intelligence profited from such difficult learning. Thus, intelligence was once again one of the best predictors of long-term learning.

Declaration of competing interest

None.

Acknowledgements

We thank Laura Dietel, Sonja Haverland, Nils Knoth, Luisa Neufeld, Lara Sokol, Tobias Steppat, Celina Stolz, Laura Wagner, and Julia Weber for their help in recruiting, data collection, and coding. We further thank Sophia Weissgerber for her help preparing and presenting the mathematical tasks in Study 1. Finally, we thank Sarah Tyrrell for proofreading the manuscript.

Funding

This research was supported by a LOEWE grant from the Hessian Ministry for Science and the Arts entitled “desirable difficulties; intrinsic cognitive motivation and performance expectancies” awarded to the co-author.

Appendix A. Example items of Study 1 (translated for this presentation, used materials in German)

Previous knowledge test:

2. Jonas takes the elevator. The equation y = -0,5x + 6 describes the relation between the elapsed time and the covered distance from Jonas. What can you deduce from this equation? Complement with which pace and from which height Jonas drives up or down.

a) Pace? _____ meter per second

b) Initial height? _____ meter

c) up/down? _____

Learning task in the reading control condition:

b) (0|0); (2|2) fits the [blue] line and has the functional equation [y = x] with the slope m= [1]; the point (22|22) ☒ is ☐ is not on this line, because [22=1·22] or [1=1].

Correct solution:

Step 1: Insert the given points (2|2) and (0|0) into the formula to compute the slope.

$m = \frac{y2-y1}{x2-x1}$ | Recap, if you enter (2|2) and (0|0) in the general formula to compute the slope $m = \frac{2-0}{2-0}$, than you will obtain $m = \frac{2}{2}$

m = 1

Step 2: This computed slope of m = 1 is inserted into the equation defining the line through the origin of y = m·x.

y = 1·x | Insert m = 1 into y = m·x

y = x

Step 3: Enter the new unknown point (22|22) with a known point on the line, e. g. (0|0) or (2|2) into the formula to compute the slope.

$m = \frac{22-0}{22-0} = \frac{22}{22} = 1$ *or* $m = \frac{22-2}{22-2} = \frac{20}{20} = 1$

Step 4: Compare, if your computed slope based on an unknown and known point in step 3 matches the previously computed slope in step 1.

m = 1 stemming from the functional equation of y = 1·x in step 1

$m = \frac{22}{22} = 1$ stemming from the computed slope of an unknown and known point in step 3

Both slopes of $m = \frac{22}{22} = 1$ and of m = 1 are equal, that is 1 = 1. Therefore, the unknown point has to fit the blue line.

alternatively:

Enter the unknown point of (22|22) into your derived functional equation from step 2 and check, whether the result is a true mathematical statement.

y = m·x | Point (22|22) has x = 22 and y = 22

22 = 1·22

22 = 22 | This is true. Thus, the unknown point (22|22) has to fit the blue line.

Learning task in the generation condition:

a) (0|0); (2|2) fits the [ ] line and has the functional equation [ ] with the slope m= [ ]; the point (22|22) ☐ is ☐ is not on this line, because [ ] or [ ].

Appendix B. Example items of Study 2 (translated for this presentation, used materials in German)

Previous knowledge test:

1. What is the functional lateralization of the brain?

Learning test in the testing condition:

2. Which sort of motor function or movement is malfunctioning due to apraxia?

Final test:

4. Wherefore is the sodium amylate test used? (Please answer the question in at most one or two sentences)

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*, 659–701. https://doi.org/10.3102/0034654316689306.

Alter, A. L., Oppenheimer, D. M., & Epley, N. (2013). Disfluency prompts analytic thinking—But not always greater accuracy: Response to. *Cognition, 128*, 252–255. https://doi.org/10.1016/j.cognition.2013.01.006.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136*, 569–576. https://doi.org/10.1037/0096-3445.136.4.569.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*, 201–210. https://doi.org/10.3758/BF03193441.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society, 2*, 59–68.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. Shimamura (Eds.). *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From learning processes to cognitive processes: Essays in honor of William K. Estes. 2. From learning processes to cognitive processes: Essays in honor of William K. Estes* (pp. 35–67).

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F., & Schwarz, N. (1994). Need for cognition: Eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben: Need for cognition: A scale measuring engagement and happiness in cognitive tasks. *Zeitschrift für Sozialpsychologie, 25*.

Bornstein, M. H., Hahn, C. S., & Wolke, D. (2013). Systems and cascades in cognitive development and academic achievement. *Child Development, 84*, 154–162. https://doi.org/10.1111/j.1467-8624.2012.01849.x.

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language, 66*, 407–415. https://doi.org/10.1016/j.jml.2011.12.009.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1563–1569. https://doi.org/10.1037/a0017021.

Carpenter, S. K., Lund, T. J., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review, 28*, 353–375. https://doi.org/10.1007/s10648-015-9311-9.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1–22. https://doi.org/10.1037/h0046743.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action. Vol. 35*. Elsevier.

Cattell, R. B., & Horn, J. L. (1978). A check on the theory of fluid and crystallized intelligence with description of new subtest designs. *Journal of Educational Measurement, 15*, 139–164. https://doi.org/10.1111/j.1745-3984.1978.tb00065.x.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354–380. https://doi.org/10.1037/0033-2909.132.3.354.

Clark, D., & Linn, M. C. (2003). Designing for knowledge integration: The impact of instructional time. *The Journal of the Learning Sciences, 12*, 451–493. https://doi.org/10.1207/S15327809JLS1204_1.

Dickhäuser, O., Schöne, C., Spinath, B., & Stiensmeier-Pelster, J. (2002). Die Skalen zum akademischen Selbstkonzept: Konstruktion und Überprüfung eines neuen Instrumentes. *Zeitschrift für differentielle und diagnostische Psychologie: ZDDP, 23*, 393–405. https://doi.org/10.1024//0170-1789.23.4.393.

Dobson, J. L., & Linderholm, T. (2015). The effect of selected "desirable difficulties" on the ability to recall anatomy information. *Anatomical Sciences Education, 8*, 395–403. https://doi.org/10.1002/ase.1489.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*, 4–58. https://doi.org/10.1177/1529100612453266.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149.

Fergusson, D. M., Horwood, L. J., & Ridder, E. M. (2005). Show me the child at seven II: Childhood intelligence and later outcomes in adolescence and young adulthood. *Journal of Child Psychology and Psychiatry, 46*, 850–858. https://doi.org/10.1111/j.1469-7610.2005.01472.x.

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review, 28*, 717–741. https://doi.org/10.1007/s10648-015-9348-9.

Gardiner, J. M., & Hampton, J. A. (1985). Semantic memory and the generation effect: Some tests of the lexical activation hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 732–741. https://doi.org/10.1037/0278-7393.11.1-4.732.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24*, 79–132. https://doi.org/10.1016/S0160-2896(97)90014-3.

Grabner, R. H., Stern, E., & Neubauer, A. C. (2006). Individual differences in chess expertise: A psychometric investigation. *Acta Psychologica, 124*, 398–420. https://doi.org/10.1016/j.actpsy.2006.07.008.

Guttman, L. (1965). A faceted definition of intelligence. *Scripta Hierosolymitana, 14*, 166–181.

Hänze, M., Schmidt-Weigand, F., & Stäudel, L. (2010). Gestufte Lernhilfen. *Innere Differenzierung in der Sekundarstufe II. Ein Praxishandbuch für Lehrer/innen* (pp. 63–73). Weinheim.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis* (2nd ed.). New York: The Guilford Press.

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: Performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology, 28*, 597–606. https://doi.org/10.1002/acp.3032.

Hirshman, E., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 484–494. https://doi.org/10.1037/0278-7393.14.3.484.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*, 253. https://doi.org/10.1037/h0023816.

Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzlestungen: Experimentell kontrollierte Weiterenwicklung eines deskriptiven Intelligenzstrukturmodells. *Diagnostica, 23*, 195–225.

Kaiser, I., Mayer, J., & Malai, D. (2018). Self-generation in the context of inquiry-based learning. *Frontiers in Psychology, 9*, 2440. https://doi.org/10.3389/fpsyg.2018.02440.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*, 23–31. https://doi.org/10.1207/S15326985EP3801_4.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93*, 579–588. https://doi.org/10.1037/0022-0663.93.3.579.

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17*, 471–479. https://doi.org/10.1080/09658210802647009.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*, 148–161. https://doi.org/10.1037/0022-3514.86.1.148.

Lehmann, J., Goussios, C., & Seufert, T. (2016). Working memory capacity and disfluency effect: An aptitude-treatment-interaction study. *Metacognition and Learning, 11*, 89–105. https://doi.org/10.1007/s11409-015-9149-z.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R (IST 2000 R). Manual (2. erweiterte und überarbeitete Aufl.)*. Göttingen: Hogrefe.

Marsh, H. W., & O'Neill, R. (1984). Self description questionnaire III: The construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement, 21*, 153–174. https://doi.org/10.1111/j.1745-3984.1984.tb00227.x.

McDaniel, M. A., & Butler, A. C. (2011). A contextual framework for understanding when difficulties are desirable. *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 175–198). .

McDaniel, M. A., Hines, R. J., & Guynn, M. J. (2002). When text difficulty benefits less-skilled readers. *Journal of Memory and Language, 46*, 544–561. https://doi.org/10.

1006/jmla.2001.2819.
McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1–43. https://doi.org/10.1207/s1532690xci1401_1.
Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*, 1474. https://doi.org/10.1037/xlm0000486.
Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H. M. (2005). Working memory and intelligence—Their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*, 61–65. https://doi.org/10.1037/0033-2909.131.1.61.
O'Neil, J. H., Spielberger, C. D., & Hansen, D. N. (1969). Effects of state anxiety and task difficulty on computer-assisted learning. *Journal of Educational Psychology, 60*, 343–350. https://doi.org/10.1037/h0028323.
Oppenheimer, D. M., & Alter, A. L. (2014). The search for moderators in disfluency research. *Applied Cognitive Psychology, 28*, 502–504. https://doi.org/10.1002/acp.3023.
Pinel, P. J., & Pauli, P. (2012). *Biopsychologie* (8. Auflage). München: Pearson Education.
Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*, 437–447. https://doi.org/10.1016/j.jml.2009.01.004.
Rapport, L. J., Brines, D. B., Theisen, M. E., & Axelrod, B. N. (1997). Full scale IQ as mediator of practice effects: The rich get richer. *The Clinical Neuropsychologist, 11*, 375–380. https://doi.org/10.1080/13854049708400466.
Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). Linking cognitive science to education: Generation and interleaving effects. *Proceedings of the twenty-seventh annual conference of the Cognitive Science Society* (pp. 1850–1855). Mahwah, NJ: Erlbaum.
Robey, A. M. (2017). *The benefits of testing: Individual differences based on student factors.* (Doctoral dissertation).
Roelle, J., & Berthold, K. (2017). Effects of incorporating retrieval into learning tasks: The complexity of the tasks matters. *Learning and Instruction, 49*, 142–156. https://doi.org/10.1016/j.learninstruc.2017.01.008.
Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence, 53*, 118–137. https://doi.org/10.1016/j.intell.2015.09.002.
Rowland, C. A. (2014, August 25). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*. https://doi.org/10.1037/a0037559 Advance online publication.
Schneider, H., Franke, B., Woisch, A., & Spangenberg, H. (2017). *Erwerb der Hochschulreife und nachschulische Übergänge von Studienberechtigten. Studienberechtigte 2015 ein halbes Jahr vor und ein halbes Jahr nach Schulabschluss* (Forum Hochschule 4|2017) Hannover: DZHW.
Schwanzer, A. (2002). *Entwicklung und Validierung eines deutschsprachigen Instruments zur Erfassung des Selbstkonzepts junger Erwachsener.* Max-Planck-Institut für Bildungsforschung.
Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592–604. https://doi.org/10.1037/0278-7393.4.6.592.
Snow, R. E. (1989). *Aptitude-treatment interaction as a framework for research on individual differences in learning*.
Snyderman, M., & Rothman, S. (1987). Survey of expert opinion on intelligence and aptitude testing. *American Psychologist, 42*, 137–144. https://doi.org/10.1037/0003-066X.42.2.137.
Spearman, C. (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology, 15*, 201–292. https://doi.org/10.2307/1412107.
Spearman, C. (1939). Thurstone's work re-worked. *Journal of Educational Psychology, 30*, 1–16. https://doi.org/10.1037/h0061267.
Spinath, B., Spinath, F. M., Harlaar, N., & Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. *Intelligence, 34*, 363–374. https://doi.org/10.1016/j.intell.2005.11.004.
Spinath, B., Stiensmeier-Pelster, J., Schöne, C., & Dickhäuser, O. (2002). *SELLMO—Skalen zur Erfassung der Lern-und Leistungsmotivation, Testmanual [SELLMO—Learning and achievement motivation assessment scales, manual].* Göttingen: Hogrefe.
Stadler, M., Becker, N., Gödker, M., Leutner, D., & Greiff, S. (2015). Complex problem solving and intelligence: A meta-analysis. *Intelligence, 53*, 92–101. https://doi.org/10.1016/j.intell.2015.09.005.
Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360–407. https://doi.org/10.1598/RRQ.21.4.1.
Stern, E. (2015). Intelligence, prior knowledge, and learning. *International Encyclopedia of the Social and Behavioral Sciences*(2nd ed.). *Vol. 12. International Encyclopedia of the Social and Behavioral Sciences* (pp. 323–328). Oxford, United Kingdom: Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.92017-8.
Stern, E. (2017). Individual differences in the learning potential of human beings. *npj Science of Learning, 2*, 2. https://doi.org/10.1038/s41539-016-0003-0.
Sternberg, R. J. (1997). The concept of intelligence and its role in lifelong learning and success. *American Psychologist, 52*, 1030–1037. https://doi.org/10.1037/0003-066X.52.10.1030.
Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35*, 401–426. https://doi.org/10.1016/j.intell.2006.09.004.
Strenze, T. (2015). *Intelligence and socioeconomic success: A study of correlations, causes and consequences.* (Doctoral dissertation).
Thurstone, L. L. (1938). *Primary mental abilities.* Chicago: University of Chicago Press.
Thurstone, L. L. (1947). *Multiple-factor analysis; a development and expansion of The Vectors of Mind.* Chicago: University of Chicago Press.
Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 607. https://doi.org/10.1037/0278-7393.5.6.607.
Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review, 27*, 247–264. https://doi.org/10.1007/s10648-015-9310-x.
Walberg, H. J., & Tsai, S. L. (1983). Matthew effects in education. *American Educational Research Journal, 20*, 359–373. https://doi.org/10.2307/1162605.
Wang, T., Ren, X., & Schweizer, K. (2017). Learning and retrieval processes predict fluid intelligence over and above working memory. *Intelligence, 61*, 29–36. https://doi.org/10.1016/j.intell.2016.12.005.
Weissgerber, S. C., & Reinhard, M. A. (2017). Is disfluency desirable for learning? *Learning and Instruction, 49*, 199–217. https://doi.org/10.1016/j.learninstruc.2017.02.004.
Wirebring, L. K., Lithner, J., Jonsson, B., Liljekvist, Y., Norqvist, M., & Nyberg, L. (2015). Learning mathematics without a suggested solution method: Durable effects on performance and brain activity. *Trends in Neuroscience and Education, 4*, 6–14. https://doi.org/10.1016/j.tine.2015.03.002.

APPENDIX D

Wenzel, K., & Reinhard, M.-A. (2021a). Does the end justify the means? Learning tests lead to more negative evaluations and to more stress experiences. *Learning and Motivation, 73*, Article 101706. https://doi.org/10.1016/j.lmot.2020.101706

This is the final article version published by Elsevier in *Learning and Motivation* available online: https://www.sciencedirect.com/science/article/abs/pii/S0023969020301995

Learning and Motivation 73 (2021) 101706

Contents lists available at ScienceDirect

Learning and Motivation

journal homepage: www.elsevier.com/locate/l&m

Does the end justify the means? Learning tests lead to more negative evaluations and to more stress experiences

Kristin Wenzel *, Marc-André Reinhard

Department of Psychology, University of Kassel, Holländische Straße 36-38, 34127 Kassel, Germany

ARTICLE INFO

Keywords:
Acute stress experiences
Negative evaluations
Test anxiety
Tests
Desirable difficulties
Situation perceptions

ABSTRACT

Although difficult learning processes like tests are beneficial for later learning outcomes, learning situations including tests or quizzes can also be perceived as acute stressors leading to more pressure, anxiety, and stress. Thus, we suppose that participants evaluate learning situations with tests, contrary to reading tasks, as more negative and experience more stress. This should be especially pronounced for learners with higher, as opposed to lower, dispositional stress or anxiety. Hence, we further predicted main effects of dispositional variables as well as interactions with the learning situation. We conducted one online study using hypothetical learning scenarios and one laboratory study using actual learning and respectively assessed dispositional stress and anxiety. Study 1 found that hypothetical learning scenarios including tests with public results and tests with private results were evaluated more negatively than re-reading control scenarios. There was also some evidence for the predicted interaction effect. In Study 2 a test in an actual learning situation was evaluated as more negative and additionally led to more acute stress and anxiety than reading. Dispositional variables were positively correlated to more negative evaluations and more stress experiences in both studies. However, there were no interactions in Study 2. Consequently, lecturers must keep in mind that learning tests can serve as acute stressors for learners, thereby resulting in negative side-effects.

1. Introduction

Educational research often focuses on ways to increase learners long-term learning outcomes and their academic achievement, for instance through applying more difficult learning situations (e.g., so called *desirable difficulties*; Bjork, 1994). However, apart from exploring beneficial effects, it is extremely valuable to test how such difficult learning situations are actually perceived. This applies to perceptions of learning situations as threats or as acute stressors, which elicit more negative evaluations as well as more experiences and feelings of anxiety, pressure, or stress. Testing such negative perceptions and experiences is important because such negative side-effects are, in turn, often linked to further negative consequences like mood disturbances, decreased motivation to learn, and reduced effort as well as persistence while learning (e.g., DeLongis, Folkman, & Lazarus, 1988; LePine, LePine, & Jackson, 2004). All in all, until now not much research was conducted to test these research questions, especially regarding tests as learning situations. Thus, the present work consists of two studies focusing on typically beneficial learning situations from a different angle: Instead of exploring effects on later long-term learning, individuals' evaluations and experiences of difficult learning situations will be tested. We want to explore how learners perceive and respond to potentially stressful learning situations, therefore taking situational factors, dispositional

* Corresponding author.
E-mail addresses: kristin.wenzel@uni-kassel.de (K. Wenzel), reinhard@psychologie.uni-kassel.de (M.-A. Reinhard).

https://doi.org/10.1016/j.lmot.2020.101706
Received 11 November 2019; Received in revised form 18 December 2020; Accepted 19 December 2020
Available online 9 January 2021

factors, as well as potential interactions into account.

We focus on desirable difficulties as intentionally hindered and challenging learning situations that were often shown to increase long-term learning and performance (e.g., Bjork, 1994; Bjork & Bjork, 1992, 2011). One of the most common, easily transferable, and robust desirable difficulties is the application of (learning or practice) tests or quizzes (also known as: *testing effect*, *retrieval practice,* or *test-enhanced learning*): Taking tests or quizzes on previously learned information—thereby actively and independently solving test questions, generating solutions, and retrieving answers—enhances long-term learning, final test performance, and retrieval of these information—even regarding curricular or complex materials (e.g., Dobson & Linderholm, 2015; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Karpicke & Aue, 2015; McDaniel, Roediger, & McDermott, 2007; recent meta-analyses: Adesope, Trevisan, & Sundararajan, 2017; Rowland, 2014; for further desirable difficulties like the strongly related *generation effect* or *disfluency*, see e.g., Bertsch, Pesta, Wiscott, & McDaniel, 2007; Diemand-Yauman, Oppenheimer, & Vaughan, 2011). Such beneficial effects of tests were found in laboratory, university, and classroom settings as well as for different learning materials like word-pairs, factual information, associations, vocabulary, longer scientific textbook paragraphs, and mathematical/statistical concepts or problems (e.g., Adesope et al., 2017; Dobson & Linderholm, 2015; Dunlosky et al., 2013; Fazio, 2017; Lim, Ng, & Wong, 2015; Lundqvist, 2019; Lyle, Bego, Hopkins, Hieb, & Ralston, 2019; Lyle & Crawford, 2011; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; Roediger & Karpicke, 2006; Rowland, 2014; Wong, Ng, Tempel, & Lim, 2019). Moreover, a wide range of learning tests and test question formats (among others triggering free recall, cued recall, or recognition) have been shown to be beneficial: These include, for instance, writing down as much of the read information as possible, multiple-choice questions, short-answer questions, fill-in-the-blank questions, completion of practice problems, finding solutions to mathematical tasks, comprehension-based questions, knowledge-based questions, application-based questions, transfer questions, and inferences (e.g., Adesope et al., 2017; Dobson & Linderholm, 2015; Dunlosky et al., 2013; Khanna, 2015; Lim et al., 2015; Lundqvist, 2019; Lyle et al., 2019; Lyle & Crawford, 2011; McDaniel et al., 2011; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Roediger & Karpicke, 2006; Rowland, 2014; Wong et al., 2019).

Theoretically, tests and desirable difficulties are supposed to trigger more retrieval practice as well as deeper (semantic) processing and encoding of information and to stimulate cognitive processes that increase understanding, retention, transfer, and also strengthen memory traces, paths, or associations (e.g., Adesope et al., 2017; Bjork, 1994; Bjork & Bjork, 1992, 2011; Karpicke, Butler, & Roediger, 2009; Roediger & Karpicke, 2006; Rowland, 2014). Desirable difficulties and tests are further assumed to lead to more analytic reasoning/thinking, more elaboration, more memory consolidation, and to the allocation of more resources concerning cognition, effort, and time (e.g., Bjork, 1994; Bjork & Bjork, 1992, 2011; Dunlosky et al., 2013; Rowland, 2014). Tests and desirable difficulties generally help to increase the meta-cognitive evaluation of learners hitherto learning success and serve as meta-cognitive cues to allocate more cognitive resources (e.g., Alter, Oppenheimer, Epley, & Eyre, 2007; Bjork, 1999; Pieger, Mengelkamp, & Bannert, 2016). Thus, learners have a more adequate and realistic view of their amount of learned information up to that point. This also reduces learners' *illusion of competence* and their general overconfidence (e.g., Bjork, 1999). Especially valuable is the effort and difficulty while being tested—the beneficial effects increase with higher effort, quality, intensity, depth, and difficulty of retrieval (e.g., Bjork & Bjork, 1992; Karpicke & Roediger, 2007; Rowland, 2014). This also applies to the format and depth of retrieval practice and learning tests: More difficult compared to easier answer formats (e.g., short answer questions vs. multiple choice questions) and deeper compared to lower questions depths (e.g., applied questions vs. factual questions) are more effective (see e.g., Maass & Pavlik, 2016).

However, such difficult learning situations were sometimes also found to pose too much additional demands as well as too much cognitive working memory load on learners, in particular regarding authentic, complex, and high element interactivity information (e.g., Clark & Linn, 2003; Roelle & Berthold, 2017; Sweller & Chandler, 1994; Van Gog & Sweller, 2015; Wenzel & Reinhard, 2019). Moreover, tests and examinations were found to moderate learners' attitudes as well as their affect—like motivation, self-efficacy, or anxiety—and were often perceived as taxing or threatening (e.g., Abouserie, 1994; Bradley et al., 2010; Crooks, 1988). Working on difficult learning tasks was also shown to increase perceived anxiety (e.g., O'Neil, Spielberger, & Hansen, 1969). Hence, specifically tests—even as learning situations—are argued to be perceived as stressful situations that lead to acute negative consequences like more negative evaluations and appraisals as well as to more acute pressure, anxiety, and stress.

1.1. Stress and anxiety in test situations

Empirically, most students experience test situations, especially final high-stake tests or examinations, as stressful and unpleasant (e.g., Beilock, 2008; Bradley et al., 2010; Hobfoll, 1989; Jamieson, Peters, Greenwood, & Altose, 2016; Sarason, 1984; Sarason & Sarason, 1990). Tests were also often correlated to feelings of threat, worry, pressure, and anxiety (e.g., Crooks, 1988; Lay, Edwards, Parker, & Endler, 1989; O'Neil et al., 1969; Stiggins, 2001). Abouserie (1994), for instance, reported that most of students' academic stress results from taking and studying for exams, from getting examination results, and from large amounts of information that must be learned. Moreover, competitive high-stake test situations, like medical school entrance examinations, were perceived as high pressure by all students, which was in turn linked to negative affective states (Leiner, Scherndl, & Ortner, 2018).

Apart from tests as (graded) examinations in university settings, even tests solely used as learning situations can be perceived as stressful situations that trigger acute negative consequences and cognitive, affective, or physiological stress responses. For instance, Hinze and Rapp (2014) found in a laboratory study—using science texts as study materials and multiple choice and open-ended application questions—that high-stake quizzes led to more anxiety and pressure than low-stake quizzes and re-reading. Low-stake quizzes were also more pressuring than re-reading. Thus, participants felt some pressure simply from taking quizzes, but higher stakes—induced by instructions stating that the monetary rewards for participants and a fictive partner were dependent of participants' quiz results—additionally led to further state anxiety. This was even independent of individual differences in trait anxiety. High-stake quizzes also negatively influenced attitudes, interests, and learning outcomes (Hinze & Rapp, 2014). In line with this,

graded quizzes (each contributing 1% to learners final grade for the respective course) in contrast to ungraded quizzes made participants feel more anxious about the course and made them less glad about having quizzes in their classes (Khanna, 2015; Khanna & Cortese, 2016).

Such learning tests may be experienced as stressful because difficult tasks are often perceived as too demanding, as overwhelming, or as consuming too many resources (e.g., Bystritsky & Kronemyer, 2014; Endler, 1997; Epel et al., 2018; Hobfoll, 1989). Besides, learners must expand cognitive resources to answer such complex questions and must actively retrieve or generate information—often under time pressure or external evaluation through peers, lecturers, or grades (e.g., Bradley et al., 2010; Jamieson et al., 2016). Additionally, tests require more effort from learners to do well, which is one aspect of a stressful situation (e.g., Epel et al., 2018). We suppose that this arises especially while working on tests as instantiations of desirable difficulties because these must be—even per definition—difficult, challenging, and demanding. They also reduce learners' illusion of competence and their overconfidence. Thus, learners achieve a more adequate but potentially unsatisfactory view of their hitherto learning progress. This is argued to lead to a perceived imbalance between the hindered tasks and learners' own capabilities and resources (see e.g., Kausar, 2010; Lazarus, 1990; McGrath, 1970). Moreover, experiencing difficulties and giving incorrect answers can feed negatively into learners' self-perceptions (e.g., Sarason & Sarason, 1990; Schunk & Gaa, 1981). Performing poorly, which often happens while working on such tasks, also further leads to experiencing stress (e.g., Sarason & Sarason, 1990; Schunk & Gaa, 1981). In line with this, learners working on difficult tasks may perceive their outcomes as less controllable, which is also an aspect of a stressor (see e.g., Epel et al., 2018).

In addition to these characteristics of the (potentially stressful) situation, dispositional variables also shape learners' appraisals, perceptions, and evaluations of different learning situations. This includes, for instance, traits stress or trait test anxiety. Trait test anxiety describes the disposition to perceive tests and test situations as stressful and threatening, including worry, emotionality, lack of confidence, and interference, which occurs especially before and during tests (e.g., Hoferichter, Raufelder, Ringeisen, Rohrmann, & Bukowski, 2016). Cassady (2004a) fittingly showed that university students with higher, as opposed to lower, dispositional cognitive test anxiety rated tests as more threatening, reported higher levels of emotionality during tests, and described relationships with helplessness attributions after tests.

The *transactional theory of stress* (e.g., Lazarus, 1990; Lazarus & Folkman, 1984, 1987) furthermore highlights the importance of individual appraisals of the situation and the importance of a conjunction of the situation and the individual. In line with this, Cassady (2004b) found that cognitive test anxiety accounted for 25 % of the variance reported in performance in a high-external evaluation pressure situation, whereas it accounted only for 12 % in a no-external evaluation pressure situation. Thus, these results showed an interaction between situational and dispositional factors.

1.2. The present research

Due to the theoretical argumentations described above and the often-observed detrimental effects of stress—like reduced motivation to learn (including reduced interest to do well, reduced effort, and reduced persistence), health problems, exhaustion, mood disturbance, emotional upset, cognitive deficits, or impaired academic performance (e.g., Brougham, Zail, Mendoza, & Miller, 2009; DeLongis et al., 1988; Hobfoll, 1989; LePine et al., 2004; Lumley & Provenzano, 2003; Struthers, Perry, & Menec, 2000)—it is important to further explore how learners actually perceive and experience learning test situations and if these actually count as acute stressors. Per definition, tests as desirable difficulties are supposed to be hindered, difficult, and complex, but still possible to overcome. We want to test linkages among learning tests, learners' immediate evaluations of such situations, and their direct stress experiences. Thereby, participants' appraisals and perceptions of different learning situations are not only manipulation checks but the foci of our work. This is relevant because researchers often argue that situations and individual perceptions of these are rarely the main focus of experiments (e.g., Edwards & Templeton, 2005). Other researchers generally highlight the importance of situation perceptions and situation classifications (e.g., Rauthmann, 2012; Rauthmann et al., 2014). Moreover, we want to test the influence of dispositional variables and potential interactions between difficult learning situations and individuals' dispositions.

Notably, not much work has been conducted regarding acute effects of tests as instantiations of desirable difficulties on (negative) situation evaluations and stress experiences. This is especially true when adding (multiple) dispositional variables like trait test anxiety and trait stress and when exploring potential interaction effects of these dispositional variables and learning situations. A lot of studies further focused on stress and anxiety inflicted by tests as examinations but not on tests as everyday learning situations.

We theoretically predict that tests, in contrast to re-reading, lead to higher negative evaluations of the learning situation, more demand, more perceived stress, and to more anxiety, especially for participants with higher levels of dispositional test anxiety, stress, or general anxiety. Supporting these assumptions, multiple studies actually showed that learning tests and quizzes lead to pressure, cognitive overload, more perceived difficulty, higher general anxiety, increased test anxiety, and more stress (e.g., Hinze & Rapp, 2014; Khanna, 2015; Van Gog & Sweller, 2015). Dispositional variables like cognitive trait test anxiety were also important for the perceptions of stress and anxiety (e.g., Cassady, 2004a, 2004b). Previous theories and studies also supported the assumption that dispositional variables and situational factors interact (e.g., Cassady, 2004b; Lazarus, 1990; Lazarus & Folkman, 1984), so that potentially stressful situations may be perceived as especially stress- and anxiety-inducing by learners that already score higher on trait stress and trait anxiety.

Although these fitting evidence exist, there are nonetheless still open questions and contrary findings. For instance, participants in other studies reported to perceive repetitive tests as learning improving and as test anxiety and negative affect reducing (e.g., Nyroos, Schéle, & Wiklund-Hörnqvist, 2016; Szpunar, Khan, & Schacter, 2013). Supposed explanations for these differing findings are that Nyroos et al. (2016) measured stress and anxiety two weeks after the test—and not immediately after or while working on it—and without a control group. Thus, the authors did not measure acute stress responses but rather long-term effects. In line with this,

Szpunar et al. (2013) measured changes in test anxiety regarding the later final test and not regarding the repetitive learning tests that were conducted before. Thus, they also assessed long-term and not immediate effects and did not focus on the learning tests as acute stressors but on later examinations as stressful situations. Because stress is argued to be a multilevel process that changes over time and that has varying effects at different time points (e.g., Epel et al., 2018; Lazarus, 1990), current ratings and momentary effects are supposed to differ from later ratings and long-term effects. Hence, although learning tests reduced anxiety regarding the final examination in the study from Szpunar et al. (2013), we argue that working on learning tests themselves still leads to more immediate negative stress responses and more stress perceptions. Moreover, the tests implemented by Nyroos et al. (2016) were worked on at home and in learners' own pace, thus, excluding potential social evaluations and resembling more self-regulated learning than learning tests applied in university courses. Most important, both studies did neither control for dispositional variables nor tested potential interactions of the learning situations and dispositional variables. Thus, the added extra value of our work and its unique contribution is that we focus specifically on learning tests as acute stressful situations and additionally test dispositional effects and interactions of the situation and the individual.

We argue that learning tests directly implemented in the university context, thus, in a real-world and regulated learning setting, lead to more acute (cognitive, affective, and physiological) stress responses like negative evaluations, more perceived anxiety, and more stress experiences. Moreover, we argue that it is extremely relevant to include measurements of learners' dispositional variables like trait stress, trait test anxiety, or general trait anxiety, and to test for interactions of the learning situations and these dispositions.

2. Study 1

Our first online Study used a learning scenario condition (two test scenario conditions and one re-reading control scenario condition) as the between-subject variable. Thereby, participants were instructed to either imagine a learning situation in which tests with private results were applied, a learning situation in which tests with (anonymous) public results were used, or a learning situation with re-reading control tasks.

We predicted that participants in both test learning scenario conditions evaluate the learning situation as more negative than participants in the re-reading control learning scenario condition (*Hypothesis 1*). Additionally, we hypothesized that dispositional variables like trait test anxiety and trait stress are positively linked to participants negative evaluation of the learning situation (*Hypothesis 2*). Moreover, we predicted interactions between these dispositional variables and the learning scenario condition: Participants with higher levels of trait test anxiety or trait stress should evaluate the test learning scenario conditions, as opposed to the re-reading control learning scenario condition, as especially negative (*Hypothesis 3*).

Further, we predicted that participants in both test learning scenario conditions experience more stress during such learning situations than participants in the re-reading control learning scenario condition (*Hypothesis 4*). Additionally, we hypothesized dispositional variables like trait test anxiety and trait stress to be positively linked to participants stress experiences (*Hypothesis 5*). Moreover, we again predicted interactions between these dispositional variables and the learning scenario condition: Participants with higher levels of trait test anxiety or trait stress should experience the test learning scenario conditions, as opposed to the re-reading control learning scenario condition, as especially stressful (*Hypothesis 6*).

2.1. Methods

In the following we report how we determined our sample size, all data exclusions, all manipulations, and all measures used in this study (cf. Simmons, Nelson, & Simonsohn, 2012).

2.1.1. Participants

We conducted an a priori power analysis using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) to calculate our required sample size: With a power of .95 and—in line with the effect sizes found by Hinze and Rapp (2014)—an assumed small to medium effect ($f = 0.20$), the power analysis for an ANCOVA with fixed effects, main effects, and interactions revealed a required sample size of $N = 390$ to detect a significant effect (alpha level of .05)—given there is one. We recruited an American online-sample consisting of 458 participants. Fifty-three of these participants were excluded because they answered at least one out of three questions testing attention incorrectly. Thus, our final sample consisted of $N = 405$ participants from MTurk ($M_{age} = 25.72$, $SD_{age} = 6.65$, range: 18–62, 48.4 % female, 97.3 % English native speakers). All participants were college or university students. Each participant was randomly assigned to one of the three learning scenario conditions: either to the learning tests with public results learning scenario ($n = 129$), the learning tests with private results learning scenario ($n = 136$), or the re-reading control learning scenario ($n = 140$) condition. Before starting, all participants had to provide their approval through reading and agreeing to an informed consent stating that they knew that their participation was completely voluntary, that they could withdraw at any time without explanations and consequences, and that they were at least 18 years old. The study was conducted in full accordance with the Ethical Guidelines of the DGPs and the APA and the project was approved by the Ethics Committee affiliated with the funding source.

2.1.2. Procedure and measures

After reading information about the study, participants reported their demographics, e.g., age, gender, native language. Thereafter, we assessed—in randomized order—dispositional variables. We measured participants trait test anxiety (PAF-E; Hoferichter et al., 2016; $\alpha = .91$) with 20 items on a four-point Likert-like scale from one (*almost never*) to four (*almost always*). Participants were thereby informed that the following statements described general feelings and thoughts in evaluation situations and that they should choose

the scale endpoint that suits them best. The scale consists of four sub-dimensions of trait test anxiety with each five items: worry (α = .80; e.g., *I worry about my results*), emotionality (α = .83; e.g., *I feel anxious*), interference (α = .91; e.g., *I easily lose my train of thoughts*), and lack of confidence (α = .88; e.g., *I am satisfied with myself*). Due the high reliability of the overall score and because the correlations among the overall score and the dependent variables were similar to the correlations among the sub-dimensions and the dependent variables, analyses will only include the overall score. We further assessed trait stress using the Perceived Stress Scale (PSS; Cohen, Kamarck, & Mermelstein, 1983), which measures the degree to which individuals perceived situations in the last month as stressful using 14 items (α = .88; e.g., *In the last month, how often have you been upset because of something that happened unexpectedly?*) on a five-point Likert-like scale from one (*never*) to five (*very often*). The Subjective Stress Scale (SSS; Reeder, Schrama, & Dirken, 1973), which uses four items (α = .88; e.g., *There is a great amount of nervous strain connected with my daily activities*) on a four-point Likert-like scale from one (*not at all*) to four (*very well*), was applied to additionally measure dispositional stress. Participants were again instructed to indicate how well the specific statements describe them in general.

We then assessed participants academic self-concept (Dickhäuser, Schöne, Spinath, & Stiensmeier-Pelster, 2002) using five items (α = .89; e.g., *My academic ability is ….*) on a seven-point Likert like scale from one (lower values) to seven (higher values) as a control variable. The academic self-concept was measured to control for potential differences between participants in the three learning scenario conditions. It was applied due to its positive correlation with learners' self-efficacy and its negative correlations with two sub-dimensions of trait test anxiety (emotionality and worry; see Dickhäuser et al., 2002). Because school-related self-efficacy is described as a tendency to approach difficult task as challenges and not as threats and was often negatively correlated to test anxiety or stress (e.g., Crooks, 1988; Hoferichter et al., 2016; Mills, Pajares, & Herron, 2006), potential differences could distort the results. The variable will only be used in the analyses if participants' academic self-concepts differ significantly between the learning scenario conditions.

Participants were then randomly assigned to one of three learning scenario conditions. They were thereby instructed to read and imagine a scenario that described learning in a college course over a whole semester that would end with an examination. In the tests with public results learning scenario condition participants were instructed to imagine that the described professor tries to increase their learning success through applying tests at the end of every session. Shortly following these sessions students receive an e-mail with a list of the matriculation numbers of all students and their respective ranked test results. In the tests with private results learning scenario condition every student receives their test results in private per e-mail and can contemplate them individually. In contrast, in the re-reading control learning scenario condition participants were instructed to imagine that the professor hands the students a summary of all relevant information to (re-)read after every session (see Appendix A for the complete scenarios). The two test learning scenario conditions were conducted—in line with Hinze and Rapp (2014)—as two slightly differing conditions that depict two forms of learning tests that are often applied in actual university settings and that vary regarding potential external or social evaluations. Thus, although tests in both scenarios are ungraded and the results presented anonymously, the imagined pressure to perform well might be higher in the tests with public results learning scenario condition because the students know that all matriculation numbers and test results are viewable for everybody else. Most import, both learning scenario conditions including tests depict realistic situations that students can actually experience in their courses and that can still be categorized as low-stake.

Following, we assessed our two dependent variables, whereas participants were instructed to answer the following question in line with perceptions and feelings they would have had if they actually were in the imagined scenarios: Firs, participants were asked to answer questions concerning their cognitive appraisals and perceptions of the learning scenario, e.g., regarding difficulty, unfairness, and feelings of anger. This resulted in an overall negative evaluation of the learning situation score using 10 items (α = .89; e.g., *Concerning the imagined scenario, … How (un)just did you find the described and imagined way of learning in the situation?,* one (*extremely unjust*) to seven (*extremely just*)) on a seven-point Likert-like scale from one (lower scores) to seven (higher scores). Higher scores thereby indicate a more negative evaluation of the situation across different domains (see Appendix A for all items and all scale endpoints). We also included three positive control items (concerning how attentive participants would be in such a situation and how helpful and interesting they would perceive the situation) that were later not analysed. These items were included so that it was not too obvious that we wanted to assess an overall negative evaluation score and to reduce the suggestiveness of the applied items. As the second dependent variable participants affective stress experience indicated by their state stress was then assessed with the Perceived Stress Questionnaire (PSQ; Levenstein et al., 1993) using 30 items (α = .95; e.g., *You feel tense*) on a four-point Likert-like scale from one (*almost never*) to four (*usually*). Participants were thereby instructed to rate the respective items in line with their stress experiences during the imagined scenarios ("*While answering the questions please imagine your feelings during the fictitious situation in class you have read. Please indicate below how you felt in the situation described by the former scenario.*").

Thereafter, we assessed variables for an unrelated study regarding desirable difficulties, negative consequence caused by learning tests, and resulting hypothetical academic cheating as well as justifications for cheating in an imagined examination (Wenzel & Reinhard, 2020). Terminating, we applied general control items and inquired if participants had really imagined the read learning scenario, if they understood the described scenario, how strongly they were able to put themselves in the scenario, and if they had ever experienced situations similar to the ones described in the scenarios. Participants received 0.60$ for their participation.

2.2. Results

Participants gender distribution, age, trait test anxiety, trait stress measured with the PSS, trait stress measured with the SSS, their academic self-concept, and the general control items did not differ between the three learning scenario conditions (all $ps \geq .160$). Thus, we did not include the academic self-concept in the analyses. All predictors were z-standardized.

Descriptive statistics of the dispositional variables—trait test anxiety and the two trait stress scales—as well as of the dependent

Table 1
Descriptive Statistics of Trait Test Anxiety, the Two Trait Stress Scales (measured with the PSS and the SSS), the Negative Evaluation of the Learning Situation, and Participants State Stress in Study 1.

Variables	M	SD	range
Trait Test Anxiety	2.29	0.64	1.00–4.00
Trait Stress measured with the PSS	2.75	0.82	1.00–5.00
Trait Stress measured with the SSS	2.22	0.85	1.00–4.00
Negative Evaluation of the Learning Situation	3.58	1.15	1.00–7.00
State Stress	2.21	0.63	1.00–4.00

Notes. $N = 405$.

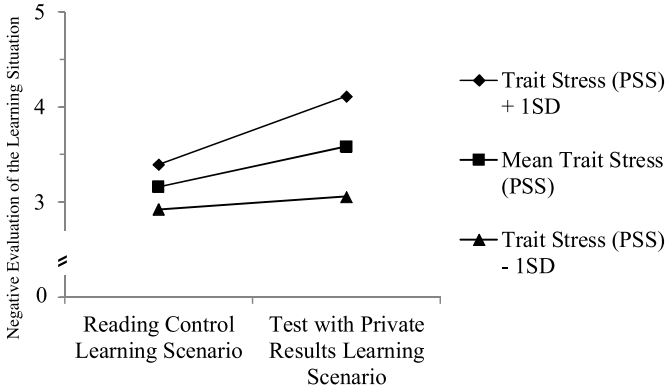


Fig. 1. The conditional effect of tests private on the negative evaluation of the learning situation for participants with relatively low (- 1SD), average, and relatively high (+ 1SD) trait stress measured with the PSS in Study 1.
Note. The learning scenario condition was dummy-coded: dummy variable 1: *tests private*, 1 = tests with private results learning scenario condition; dummy variable 2: *tests public*, 1 = tests with public results learning scenario condition; reference category: 0 = re-reading control learning scenario condition.

variables—negative evaluation of the learning situation and participant state stress—are depicted in Table 1. Notably, trait test anxiety was significantly correlated to trait stress measured with the PSS and to trait stress measured with the SSS ($r = .65$, $r = .60$, respectively; both $ps < .001$). The two trait stress scales were also significantly correlated to each other ($r = .67$, $p < .001$).

2.2.1. Negative evaluation of the learning situation

To test Hypothesis 1, we conducted an ANOVA to compare participants negative evaluation of the learning situation in the three learning scenario conditions: There was a significant main effect of the learning scenario condition, $F(2,402) = 22.16$, $p < .001$, $\eta_p = .10$. Results (and subsequent pairwise comparisons) indicated that participants in the tests with public results learning scenario condition rated the learning situation as significantly more negative ($M_{public} = 4.04$, $SD_{public} = 1.06$) than participants in the tests with private results learning scenario condition ($M_{private} = 3.58$, $SD_{private} = 1.11$) and participants in the re-reading control learning scenario condition ($M_{re\text{-}reading} = 3.15$, $SD_{re\text{-}reading} = 1.20$; both $ps < .001$). A further pairwise comparison showed that the tests with private results learning scenario condition was evaluated as significantly more negative than the re-reading control learning scenario condition ($p = .001$).

Concerning Hypothesis 2, we firstly regarded the correlations of the negative evaluation of the learning situation and the three dispositional variables trait test anxiety, trait stress measured with the PSS, and trait stress measured with the SSS ($r = .28$, $p < .001$; $r = .33$, $p < .001$; $r = .31$, $p < .001$; respectively). To test this hypothesis in more detail and to test which predictor remains robust after controlling for the other variables, we then conducted a regression analysis including the three dispositional variables as predictors for participants negative evaluation of the learning situation. R for this regression analysis was significantly different from zero, $F(3,401) = 19.57$, $R^2 = .128$, $R^2_{adj} = .121$, $p < .001$. Trait test anxiety was not a significant predictor, $t(401) = 1.23$, $B = 0.09$, $SD = 0.07$, $\beta = .079$, $p = .221$. Trait stress measured with the PSS was, however, significant, $t(401) = 2.70$, $B = 0.22$, $SD = 0.08$, $\beta = .187$, $p = .007$. Trait stress measured with the SSS was also significant, $t(401) = 2.13$, $B = 0.16$, $SD = 0.08$, $\beta = .140$, $p = .034$.

To test Hypothesis 3, we conducted two dummy variables for the learning scenario condition (dummy variable 1: *tests private*, 1 = tests with private results learning scenario condition; dummy variable 2: *tests public*, 1 = tests with public results learning scenario condition; reference category: 0 = re-reading control learning scenario condition). Using SPSS and PROCESS (model 1; Hayes, 2018) we then ran a regression analysis predicting the negative evaluation of the learning situation using tests private, tests public, trait stress measured with the PSS, as well as the interactions of tests private and tests public with trait stress measured with the PSS. To keep the regression analysis as simple as possible, we only used one dispositional variable and the respective interactions instead of using all three dispositional variables and all six potential interactions simultaneously. We chose trait stress measured with the PSS as the dispositional variable because it was the descriptively strongest dispositional variable in the regression analysis presented above. R for this regression analysis was significantly different from zero, $F(5,399) = 21.57$, $R^2 = .213$, $R^2_{adj} = .203$, $p < .001$. Tests private was a

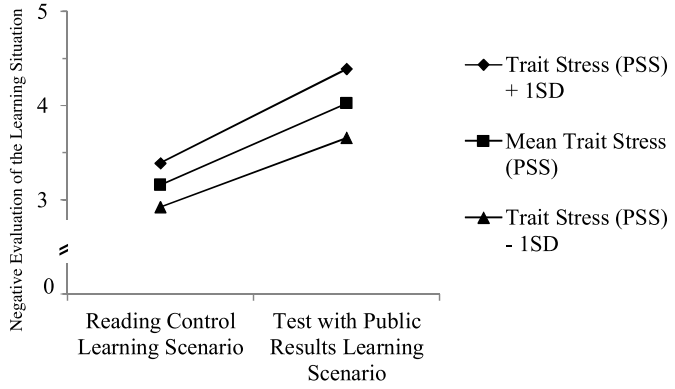


Fig. 2. The conditional effect of tests public on the negative evaluation of the learning situation for participants with relatively low (- 1SD), average, and relatively high (+ 1SD) trait stress measured with the PSS in Study 1.
Note. The learning scenario condition was dummy-coded: dummy variable 1: *tests private*, 1 = tests with private results learning scenario condition; dummy variable 2: *tests public*, 1 = tests with public results learning scenario condition; reference category: 0 = re-reading control learning scenario condition.

significant predictor, $t(399) = 3.43$, $B = 0.43$, $SD = 0.12$, $\beta = .174$, $p = .001$. Tests public was also significant, $t(399) = 6.89$, $B = 0.87$, $SD = 0.13$, $\beta = .350$, $p < .001$. Trait stress measured with the PSS was significant as well, $t(399) = 2.71$, $B = 0.24$, $SD = 0.09$, $\beta = .204$, $p = .007$. The interaction of tests private and trait stress measured with the PSS was significant, $t(399) = 2.32$, $B = 0.29$, $SD = 0.13$, $\beta = .142$, $p = .021$. A closer look at the conditional effects of this ordinal interaction (see Fig. 1) revealed that there was no significant effect of tests private on the negative evaluation of the learning situation for participants with rather low trait stress (trait stress measured with the PSS 1SD below mean), $t(399) = 0.75$, $B = 0.13$, $SE = 0.18$, $p = .453$. However, participants with average trait stress in the tests with private results learning scenario condition evaluated the learning situation as more negative than participants in the re-reading control learning scenario condition, $t(399) = 3.43$, $B = 0.43$, $SE = 0.12$, $p = .001$. This effect was especially strong for participants with higher trait stress (trait stress measured with the PSS 1SD above mean), $t(399) = 4.02$, $B = 0.72$, $SE = 0.18$, $p < .001$. The interaction of tests public and trait stress measured with the PSS was, however, not significant, $t(399) = 1.06$, $B = 0.13$, $SD = 0.12$, $\beta = .066$, $p = .292$[1] (see Fig. 2; see Appendix C and Fig. C1 for another figure depicting the significant interaction of tests private and trait stress and the not significant interaction of tests public and trait stress in one figure).

Notably, these findings supported Hypothesis 1 showing a main effect of the learning scenario condition on the negative evaluation of the learning situation, insofar as that learning scenarios including tests were evaluated as more negative than the reading control scenario. Hypothesis 2 was also supported by the found positive correlations among participants' trait variables and the negative evaluation of the learning situation. Hypothesis 3, assuming an interaction between the learning scenario condition and participants trait variables, was only partly supported by our data due to the significant interaction between tests private and trait stress but the not significant interaction between tests public and trait stress. Interestingly, low trait stress buffered the negative effect of the tests with private results learning scenario condition on participants negative evaluation of the situation. Tests public, however, increased the negative evaluation of the learning situation for every participant, independent of their trait stress.

2.2.2. Stress experiences

To test Hypothesis 4, we conducted an ANOVA to compare participants stress experiences indicated by their state stress in the three learning scenario conditions: There was no significant main effect of the learning scenario condition on state stress, $F(2, 402) = 1.73$, $p = .179$, $\eta_p = 0.01$. Thus, there were no significant differences among participants in the tests with public results learning scenario condition ($M_{public} = 2.29$, $SD_{public} = 0.64$), participants in the tests with private results learning scenario condition ($M_{private} = 2.15$, $SD_{private} = 0.63$), and participants in the re-reading control learning scenario condition ($M_{re\text{-}reading} = 2.20$, $SD_{re\text{-}reading} = 0.60$).

Concerning Hypothesis 5, we firstly regarded the correlations of state stress and the three dispositional variables trait test anxiety, trait stress measured with the PSS, and trait stress measured with the SSS ($r = .56$, $p < .001$; $r = .52$, $p < .001$; $r = .54$, $p < .001$; respectively). To test this hypothesis in more detail and to test which predictor remains robust after controlling for the other variables, we then conducted a regression analysis including the three dispositional variables as predictors for participants state stress. *R* for this regression analysis was significantly different from zero, $F(3,401) = 85.60$, $R^2 = .390$, $R^2_{adj} = .386$, $p < .001$. Trait test anxiety was a significant predictor, $t(401) = 5.70$, $B = 0.19$, $SD = 0.03$, $\beta = .306$, $p < .001$. Trait stress measured with the PSS was also significant, $t(401) = 2.58$, $B = 0.09$, $SD = 0.04$, $\beta = .149$, $p = .010$. Trait stress measured with the SSS was significant as well, $t(401) = 4.75$, $B = 0.16$, $SD = 0.03$, $\beta = .261$, $p < .001$.

To test Hypothesis 6, we then conducted a regression analysis using SPSS and PROCESS (model 1; Hayes, 2018) predicting state

[1] Using trait stress measured with the SSS instead of trait stress measured with the PSS leads to mirroring findings. Tests private, tests public, trait stress measured with the SSS, as well as the interaction of tests private and trait stress measured with the SSS were significant. The interaction thereby followed the same pattern as the interaction above. Using trait test anxiety instead of trait stress measured with the PSS leads to partially similar results, finding tests private, tests public, and trait test anxiety to be significant. Both interactions were, however, not significant.

stress using tests private, tests public, trait test anxiety, as well as the interactions of tests private and tests public with trait test anxiety. We chose trait test anxiety as the dispositional variable because it was the descriptively strongest dispositional variable in the regression analysis presented above. R for this regression analysis was significantly different from zero, $F(5,399) = 38.08$, $R^2 = .323$, $R^2_{adj} = .315$, $p < .001$. Tests private was not a significant predictor, $t(399) = -0.74$, $B = -0.05$, $SD = 0.06$, $\beta = -.035$, $p = .461$. Tests public was also not significant, $t(399) = 1.62$, $B = 0.10$, $SD = 0.06$, $\beta = .076$, $p = .106$. Trait test anxiety was a significant predictor, $t(399) = 7.54$, $B = 0.36$, $SD = 0.05$, $\beta = .579$, $p < .001$. The interaction of tests private and trait test anxiety was not significant, $t(399) = -0.56$, $B = -0.04$, $SD = 0.07$, $\beta = -.034$, $p = .573$. The interaction of tests public and trait test anxiety was not significant as well, $t(399) = 0.03$, $B = 0.002$, $SD = 0.06$, $\beta = .002$, $p = .973$.[2]

Notably, these findings did neither support Hypothesis 4 nor Hypothesis 6 because the learning scenario condition had no effect on participants stress experiences and there were no significant interactions between the learning scenario conditions and participants trait variables. Hypothesis 5, predicting positive correlations among participants' trait variables and their stress experiences, was supported. Neither further controlling for participants gender nor for their academic self-concept changed any of the results of our first study.

2.3. Discussion

This study was conducted to test if learning tests—in form of imagined learning scenarios—lead to a more negative evaluation of the learning situation and to more stress experiences than a re-reading learning scenario. Additionally, we checked for effects of multiple dispositional variables and potential interactions of these and the learning scenario condition.

As expected, both learning scenarios including tests significantly led to more negative evaluations than the learning scenario including re-reading. Scenarios with tests with public results were evaluated the most negative. Interestingly, this occurred although the learning tests were only presented in a very short hypothetical scenario, were ungraded, introduced as learning improving, and without consequences for participants actual or imagined courses. Both scenarios including tests can therefore be described as low-stakes even though pressure to perform well should have been slightly higher in the learning tests with public results learning scenario condition. Thus, effects in real classrooms—including real learning materials, actual peers, or incentives to do well—are probably even more pronounced. The results also remained robust when adding dispositional variables like trait stress and trait test anxiety, which were also positively linked to participants negative evaluation of the learning situation. Notably, we found a significant interaction between tests private and trait stress. We had previously predicted that all participants should evaluate the learning scenarios with tests more negatively than the re-reading control scenario and that these negative effects should be especially strong for participants with higher trait stress. However, the negative evaluation of the learning situation did not differ between participants in the private results learning scenario condition and participants in the re-reading control learning scenario condition for participants with lower trait stress. Thus, lower trait stress was beneficial and buffered the negative effect of the tests with private results learning scenario condition on participants evaluation of the learning situation. The negative effect of the learning scenario with private tests was, however, significant for participants with average trait stress, and even descriptively higher for participants with higher trait stress. There was, in contrast, no significant interaction between tests public and the dispositional variables. Hence, lower trait stress was apparently not able to buffer the negative effect of the tests with public results learning scenario condition, which was perceived as the most negative of the three learning scenario conditions.

Unexpectedly, there was no effect of the learning scenario condition on participants stress experiences. Only trait test anxiety and the trait stress scales were positively linked to participants stress experiences. There were also no interactions between the learning scenario condition and these dispositional variables. We suppose that these non-significant results were due to our operationalization: The hypothetical scenarios were probably not strong enough to elicit actual affective experiences of stress. In line with this, our learning scenarios were rather short and included not many details. Thus, we were not able to control which courses, materials, and complexity or difficulty levels participants actually imagined or how strongly (if at all) participants were even able to imagine the presented scenarios and the described learning. The same applies to the predicted interaction effects that were—with one exception concerning negative evaluations—not found. Hence, although scenarios are important for first evidence, effects in actual classroom or laboratory settings are advantageous for testing learners' actual acute experiences in learning situations. Moreover, we argued in the present research section that Nyroos et al. (2016) found no effects of tests on stress and anxiety because learners rated their stress experiences two weeks after the test and were simply not able to correctly recall their (past) feelings and experiences. The same could apply to our operationalization, insofar as that participants might have struggled to imagine acute stress they would have experienced when indulging in multiple (hypothetical) learning tests over a whole (hypothetical) semester. Thus, it would be valuable to instead focus on stress experiences directly after one distinct learning situation—to focus on immediate and acute effects—as well as to apply actual and realistic learning materials instead of only imagined learning materials that were not even described in detail. Moreover, future work could also add even more dependent variables to cover all potential aspects of stress experiences.

[2] Using trait stress measured with the PSS or trait stress measured with the SSS instead of trait test anxiety leads to mirroring findings. In each case only the dispositional variable would be a significant predictor, but neither tests private, tests public, nor the interactions of the dummy-coded learning scenario condition and the respective dispositional variable.

3. Study 2

In line with the argumentations and discussion stated directly above, Study 2 included actual learning in a laboratory and was conducted to resemble normal university courses as closely as possible. The between-subject variable was the learning condition (one single test or one reading control condition). To increase our understanding of stress experiences during learning situations, we additionally implemented further dependent variables. Thus, we now measure the (cognitive) negative evaluation of the learning situation and three varying indicators for stress experiences, two affective indicators—state stress and general state anxiety—as well one physiological indicator—participants pulse—directly after the learning phase.

In line with Study 1, we predicted that participants in the test condition, contrary to participants in the reading control condition, evaluate the learning situation as more negative (*Hypothesis 1*). Additionally, we hypothesized that dispositional variables like trait test anxiety, trait stress, and general trait anxiety are positively linked to participants negative evaluation of the learning situation (*Hypothesis 2*). Moreover, we predicted interactions between these dispositional variables and the learning condition: Participants with higher levels of trait test anxiety, trait stress, or general trait anxiety should evaluate the test condition, as opposed to the reading control condition, as especially negative (*Hypothesis 3*).

Furthermore, we hypothesized that participants in the test condition, contrary to participants in the reading control condition, experience more stress. Indicators for such stress experiences are higher self-reported state stress (*Hypothesis 4a*), more general state anxiety (*Hypothesis 4b*), and a higher pulse (*Hypothesis 4c*). Additionally, we predicted that dispositional variables like trait test anxiety, trait stress, and general trait anxiety are positively linked to stress experiences indicated by state stress (*Hypothesis 5a*), general state anxiety (*Hypothesis 5b*), and pulse (*Hypothesis 5c*). Moreover, we hypothesized interactions between these dispositional variables and the learning condition: Participants with higher levels of trait test anxiety, trait stress, or general trait anxiety in the test condition, as opposed to the reading control condition, should have especially high stress experiences indicated by state stress (*Hypothesis 6a*), general state anxiety (*Hypothesis 6b*), and pulse (*Hypothesis 6c*).

3.1. Methods

In the following we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures assessed and used in this study (cf. Simmons et al., 2012).

3.1.1. Participants

We again conducted an a priori power analysis using G*Power (Faul et al., 2009) to calculate our required sample size: With a power of .85[3] and—in line with the effect sizes of our first study—an assumed medium effect ($f = 0.30$), the power analysis for an ANCOVA with fixed effects, main effects, and interactions revealed a required sample size of $N = 102$ to detect a significant effect (alpha level of .05)—given there is one. Thus, we recruited a sample consisting of 108 participants ($M_{age} = 23.26$, $SD_{age} = 5.24$, range: 18–54, 73.1 % female, 89.8 % German native speakers, 97.2 % university students). Each participant was randomly assigned to one of the two learning conditions: either the test ($n = 53$) or the reading control ($n = 55$) condition. Before starting, all had to provide their approval through reading and agreeing to a written informed consent. The study was conducted in full accordance with the Ethical Guidelines of the DGPs and the APA and the project was approved by the Ethics Committee affiliated with the funding source.

3.1.2. Procedure and measures

Up to eight participants could simultaneously take part; at least five were present in every test condition. The laboratory setting consisted of rows of tables, so that participants faced the experimenter in the front like students face lecturers in university courses.

In the beginning, participants reported their demographics, e.g., age, sex, native language. Thereafter, we assessed—in randomized order—dispositional variables. These were chosen following the trait variables applied in Study 1. Thus, we used similar German scales and further included another measure—general trait anxiety—to broaden the scope of the assessed dispositional variables. We measured participants trait test anxiety using the German version of the scale from Study 1 (PAF; Hodapp, Rohrmann, & Ringeisen, 2011; $\alpha = .88$; worry: $\alpha = .84$; emotionality: $\alpha = .86$; interference: $\alpha = .86$; lack of confidence: $\alpha = .86$). Our analyses again use the overall score. Trait stress was assessed using the screening scale of the Trier Inventory for Chronic Stress (TICS; Schulz, Schlotz, & Becker, 2004), which includes 12 items assessing the amount of different stressful events in the last three months ($\alpha = 88$; e.g., concerning worry or pressure to perform) on a five-point Likert-like scale from one (*never*) to five (*very often*). Additionally, general trait anxiety was measured with the German trait short version of the State Trait Anxiety Inventory (STAI-T; Laux, Glanzmann, Schaffner, & Spielberger, 1981) using 10 items ($\alpha = .80$; e.g., *I get tired quickly*) on an eight-point Likert-like scale from one (*almost never*) to eight (*almost always*). Before rating any one of these scales, participants were instructed to indicate how well the specific statements describe them in general.

Then the learning phase started. We first applied an initial study opportunity in which each participant read basic information on mathematical equations, linear functions, and regression analyses. They were also presented with some short example solutions regarding the described information. The applied materials should not be new, unknown, or too demanding for participants because

[3] The power level was slightly smaller than the power level applied in Study 1 to reduce the number of participants needed to detect significant effects. This was done due to the higher difficulty of recruiting participants for a laboratory study compared to recruiting participants for online studies.

the included information and concepts were normally already learned during their school times—thus, the underlying concepts can be classified as basic. The materials should, however, not be repeated or used regularly during participants university courses. All in all, the learned information consists of relevant, realistic, and curricular school/university learning materials that should be challenging but not too difficult or overwhelming. Moreover, most students have to learn mathematical or statistical information during school or university (often as basis for further information) and previous studies also often used mathematical information when testing the effects of tests and desirable difficulties (see e.g., Bertsch et al., 2007; Fazio, 2017; Lim et al., 2015; Lyle et al., 2019; Lyle & Crawford, 2011; Lundqvist, 2019; Wirebring et al., 2015; Wong et al., 2019). Following this first learning phase, we then assessed individuals' task-specific self-concept (see Dickhäuser & Reinhard, 2006; Dickhäuser et al., 2002). The scale included five German items ($\alpha = 97$; e.g., *My ability for such tasks is …*) on a seven-point Likert-like scale from one (lower values) to seven (higher values). The items were similar to the ones assessing participants academic ability in Study 1 but referred to the specific learning materials and not to participants general academic self-concept. As before, this trait variable was only assessed to control for potential differences between participants in the two learning conditions and will only be included in later analyses if the mean scores differ significantly between participants in both learning conditions.

Then, as a second learning phase, all participants worked simultaneously for 10 min on a learning task that consisted of a work sheet presenting arithmetic problems, e.g., identifying a line based on two given points in a graph (see Appendix B). They were instructed to learn as intensely as they would normally learn in their actual courses and to do their best. Participants were randomly distributed to one of the two manipulated learning conditions, either the reading control condition or the test condition. In the reading control condition participants were instructed to read, understand, and memorize the already answered tasks and the needed solution steps. Thus, participants were able to restudy the information, explanations, and solutions steps that they had read in the first learning phase directly before. In contrast, participants in the test condition had to take a learning test on the previously read and studied materials: They had 6 min to think about the presented problems, to retrieve the learned information and explanations, and to recall the needed solutions steps presented in the initial study phase to then actively answer and solve the test questions. The test included mostly short-answer and fill-in-the-blank questions. The experimenter then vocally tested their answers; in randomized order every participant had to answer two of the questions aloud and received short feedback regarding the accuracy of their answers. This increased the similarity of the test condition to everyday university courses in which students must also answer (test) questions and retrieve previously learned information while being surrounded by peers. Additionally, every participant thereby received short feedback concerning the learning test answers—individually correcting every learning test could be too time-consuming for lecturers. Most important, the learning test was a low-stake situation because participants knew that the test was ungraded, not related to their actual courses, did not influence their monetary reward for participation, and was conducted as a learning situation and not as an examination.

Almost at the end of this learning phase—after 8 of the 10 min—the experimenter then instructed participants to measure their pulse using a pulse oximeter. Pulse thereby served as an indicator for a physiological stress experience. Unfortunately, however, the pulse assessment was seemingly not as easy as we supposed because three participants reported impossible scores (0, 2.5, and 4). It is thus possible that other scores were wrong as well but were simply not detected because they seemed apparently correct. Hence, later analyses testing the hypotheses concerning participants pulse could be contorted and can only be interpreted with caution—or could be even completely uninterpretable. We therefore relocated all analyses concerning this dependent variable as well as further discussions on the issue with its interpretation to the Appendix (see Appendix C).[4]

Immediately after the completed second learning phase, the negative evaluation of the learning situation was measured as a dependent variable: Participants were instructed to rate their evaluations and perceptions of and during this second learning phase using the German version of the items applied in Study 1 ($\alpha = .89$; see Appendix A). Next, we measured participants affective stress experiences using the other two indicators apart from pulse: self-reported state stress and self-reported general state anxiety. Participants state stress was measured with a German version of the Perceived Stress Questionnaire (PSQ; Fliege, Rose, Arck, Levenstein, & Klapp, 2001; based on Levenstein et al., 1993), which includes 20 items (e.g., *You feel tense*). Participants were again instructed to refer their ratings to the second learning phase. In line with adding this introduction, we also discussed about changing the endpoints of the scale to make it even clearer that the items refer to participants' immediate situational experiences. Unfortunately, we erroneously only assessed 10 items ($\alpha = .88$) on the normal four-point Likert-like scale from one (*almost never*) to four (*usually*). The other 10 items ($\alpha = .92$) were assessed on a seven-point Likert-like scale from one (*totally disagree*) to seven (*totally agree*). We z-standardized both halves of the scale and then conducted a mean score of these two z-standardized variables ($\alpha = .94$). Participants general state anxiety was assessed using the German state short version of the State Trait Anxiety Inventory (STAI-S; Laux et al., 1981) with 10 items ($\alpha = .90$; e.g., *I am calm*) on a seven-point Likert-like scale from one (*almost never*) to seven (*almost always*). Again, participants were instructed to respond to these items in regard to the second learning phase experienced directly before.

Finally, participants were asked if they had participated in a similar study or if they knew the applied learning materials. They got 5 Euro—or in case of psychology students course credit—as a reward for their participation.

3.2. Results

Participants gender distribution, age, trait test anxiety, trait stress, and general trait anxiety did not differ between the two learning

[4] We want to thank an anonymous Reviewer for this suggestion.

Table 2
Descriptive Statistics of Trait Test Anxiety, Trait Stress, General Trait Anxiety, the Negative Evaluation of the Learning Situation, and Participants Stress Experiences indicated by State Stress and General State Anxiety in Study 2.

Variables	M	SD	range
Trait Test Anxiety	2.29	0.46	1.20–3.75
Trait Stress	3.01	0.66	1.58–4.75
General Trait Anxiety	4.14	1.05	1.50–7.10
Negative Evaluation of the Learning Situation	3.29	1.12	1.20–6.30
State Stress	0.00	0.94	−1.80 to 1.96
General State Anxiety	3.35	1.15	1.30–6.30

Notes. $N = 108$. State Stress was z-standardized.

conditions (all $ps \geq .153$). The task-specific self-concept, however, differed significantly between the two learning conditions, $M_{test} = 4.52$, $SD_{test} = 1.27$, $M_{reading} = 3.89$, $SD_{reading} = 1.65$, $t(106) = 2.22$, $p = .029$. Participants in the test condition had a significantly higher task-specific self-concept than participants in the reading control condition. Thus, our analyses will include this variable. All predictors were z-standardized.

Descriptive statistics of the dispositional variables—trait test anxiety, trait stress, and general trait anxiety—and the dependent variables—negative evaluation of the learning situation, state stress, and general state anxiety—are depicted in Table 2. Trait test anxiety was significantly and positively correlated to trait stress and to general trait anxiety ($r = .50$, $r = .59$, respectively; both $ps < .001$). Trait stress and general trait anxiety were also correlated to each other ($r = .53$, $p < .001$).

3.2.1. Negative evaluation of the learning situation

To test Hypothesis 1, we conducted a *t*-test to compare the average negative evaluation of the learning situation of participants in both learning conditions: $M_{test} = 3.43$, $SD_{test} = 1.12$, $M_{reading} = 3.15$, $SD_{reading} = 1.12$, $t(106) = 1.31$, $p = .192$. Unexpectedly, there was no significant difference. Due to the above-mentioned significant difference of the task-specific self-concept between participants in both learning conditions, we further conducted a regression analysis predicting the negative evaluation of the learning situation through the learning condition (0 = reading control condition, 1 = test condition) and the task-specific self-concept. *R* for this regression analysis was significantly different from zero, $F(2,105) = 25.14$, $R^2 = .324$, $R^2_{adj} = .311$, $p < .001$. The learning condition was now a significant predictor, $t(105) = 3.00$, $B = 0.55$, $SD = 0.18$, $\beta = .246$, $p = .003$. The task-specific self-concept was also significant, $t(105) = -6.91$, $B = -0.64$, $SD = 0.09$, $\beta = -.568$, $p < .001$.

Concerning Hypothesis 2, we firstly regarded the correlations of the negative evaluation of the learning situation and the three dispositional variables trait test anxiety, general trait anxiety, and trait stress ($r = .19$, $p = .050$; $r = .19$, $p = .052$; $r = .06$, $p = .551$; respectively). To test our hypothesis more thoroughly and to test which (or if any) predictor remains robust after controlling for the other variables, we conducted a regression analysis including all three dispositional variables as predictors for participants negative evaluation of the learning situation. *R* for this regression analysis was not significantly different from zero, $F(3, 104) = 1.85$, $R^2 = .051$, $R^2_{adj} = .023$, $p = .142$. None of the predictors was significant: Neither trait test anxiety, $t(104) = 1.20$, $B = 0.17$, $SD = 0.14$, $\beta = 149$, $p = .231$, nor general trait anxiety, $t(104) = 1.20$, $B = 0.17$, $SD = 0.14$, $\beta = .150$, $p = .232$, nor trait stress, $t(104) = -0.82$, $B = -0.11$, $SD = 0.13$, $\beta = -.096$, $p = .414$.

To test Hypothesis 3, we then conducted another regression analysis using SPSS and PROCESS (model 1; Hayes, 2018) to predict the negative evaluation of the learning situation using the learning condition, task-specific self-concept, trait test anxiety, as well as the interaction of the learning condition and trait test anxiety. We chose trait test anxiety as the dispositional variable because the correlation of trait test anxiety and the negative evaluation of the learning situation was—contrary to the correlations of the other two dispositional variables—marginally significant. *R* for this regression analysis was significantly different from zero, $F(4,103) = 13.79$, $R^2 = .349$, $R^2_{adj} = .323$, $p < .001$. The learning condition was a significant predictor, $t(103) = 2.93$, $B = 0.54$, $SD = 0.18$, $\beta = .239$, $p = .004$. The specific self-concept was also significant, $t(103) = -6.83$, $B = -0.42$, $SD = 0.06$, $\beta = -.566$, $p < .001$. Trait test anxiety was not significant, $t(103) = -0.10$, $B = -0.01$, $SD = 0.12$, $\beta = -.011$, $p = .920$. The interaction of the learning condition and trait test anxiety was also not significant, $t(103) = 1.57$, $B = -0.29$, $SD = 0.18$, $\beta = .165$, $p = .119$.[5] Notably, these findings supported Hypothesis 1 predicting that the test condition would be evaluated as more negative than the reading control condition. However, neither Hypothesis 2 nor Hypothesis 3 regarding the predicted positive correlations among participants' trait variables and the negative evaluation of the learning situation as well as regarding the predicted interactions between the learning situation and trait variables were supported.

3.2.1.1. Stress experiences indicated by state stress. To test Hypothesis 4a, we conducted a *t*-test to compare participants stress experiences indicated by their average state stress in both learning conditions: $M_{test} = 0.15$, $SD_{test} = 0.95$, $M_{reading} = -0.15$, $SD_{reading} = 0.90$, $t(106) = 1.70$, $p = .093$. Unexpectedly, there was no significant difference. Again, due to the above-mentioned significant difference of the task-specific self-concept between participants in both learning conditions, we further conducted a regression analysis predicting

[5] Using general trait anxiety or trait stress instead of trait test anxiety leads to mirroring findings. In each case, the learning condition and the task-specific self-concept were significant predictors but neither the dispositional variable nor the interaction of the learning condition and the dispositional variable were significant.

state stress through the learning condition (0 = reading control condition, 1 = test condition) and the task-specific self-concept. *R* for this regression analysis was significantly different from zero, $F(2,105) = 22.73$, $R^2 = .302$, $R^2_{adj} = .289$, $p < .001$. The learning condition was now a significant predictor, $t(105) = 3.30$, $B = 0.51$, $SD = 0.16$, $\beta = .276$, $p = .001$. The task-specific self-concept was also significant, $t(105) = -6.44$, $B = -0.50$, $SD = 0.08$, $\beta = -.537$, $p < .001$.

Concerning Hypothesis 5a, we firstly regarded the correlations of state stress and the three dispositional variables trait test anxiety, general trait anxiety, and trait stress ($r = .42$, $p < .001$; $r = .32$, $p = .001$; $r = .16$, $p = .104$; respectively). To test this hypothesis more thoroughly and to test which predictor remains robust when controlling for the other variables, we then conducted a regression analysis including all three dispositional variables as predictors for state stress. *R* for this regression analysis was significantly different from zero, $F(3,104) = 8.35$, $R^2 = .194$, $R^2_{adj} = .171$, $p < .001$. Trait test anxiety was a significant predictor, $t(104) = 3.42$, $B = 0.36$, $SD = 0.11$, $\beta = .388$, $p = .001$. In contrast, general trait anxiety, $t(104) = 1.35$, $B = 0.15$, $SD = 0.11$, $\beta = .156$, $p = .180$, and trait stress, $t(104) = -1.12$, $B = -0.11$, $SD = 0.10$, $\beta = -.120$, $p = .267$, were not significant.

To test Hypothesis 6a, we conducted another regression analysis using SPSS and PROCESS (model 1; Hayes, 2018) to predict stress experiences indicated by state stress using the learning condition, task-specific self-concept, trait test anxiety, as well as the interaction of the learning condition and trait test anxiety. We chose trait test anxiety as the dispositional variable because it was the only significant dispositional variable in the regression analysis presented above. *R* for this regression analysis was significantly different from zero, $F(4,103) = 18.72$, $R^2 = .421$, $R^2_{adj} = .398$, $p < .001$. The learning condition was a significant predictor, $t(103) = 3.21$, $B = 0.46$, $SD = 0.14$, $\beta = .247$, $p = .002$. The specific self-concept was also significant, $t(103) = -6.31$, $B = -0.46$, $SD = 0.07$, $\beta = -.493$, $p < .001$. Trait test anxiety was also significant, $t(103) = 2.60$, $B = 0.24$, $SD = 0.09$, $\beta = .260$, $p = .011$. The interaction of the learning condition and trait test anxiety was, however, not significant, $t(103) = 1.20$, $B = 0.17$, $SD = 0.14$, $\beta = .118$, $p = .233$.[6] These findings supported Hypothesis 4a showing that the test condition increased participants state stress compared to participants in the reading control condition and partly supported Hypothesis 5a regarding positive correlations among trait variables and state stress. Hypothesis 6a regarding the predicted interaction of the situation and the individual was, however, not supported.

3.2.1.2. Stress experiences indicated by general state anxiety. To test Hypothesis 4b, we conducted a *t*-test to compare participants stress experiences indicated by their average general state anxiety in both learning conditions: $M_{test} = 3.59$, $SD_{test} = 1.17$, $M_{reading} = 3.12$, $SD_{reading} = 1.10$, $t(106) = 2.15$, $p = .034$. We again additionally conducted a regression analysis predicting general state anxiety through the learning condition (0 = reading control condition, 1 = test condition) and the task-specific self-concept. *R* for this regression analysis was significantly different from zero, $F(2,105) = 9.66$, $R^2 = .155$, $R^2_{adj} = .139$, $p < .001$. The learning condition remained a significant predictor, $t(105) = 3.02$, $B = 0.64$, $SD = 0.21$, $\beta = .277$, $p = .003$. The task-specific self-concept was also significant, $t(105) = -3.76$, $B = -0.40$, $SD = 0.11$, $\beta = -.345$, $p < .001$.

Concerning Hypothesis 5b, we firstly regarded the correlations of general state anxiety and the three dispositional variables trait test anxiety, general trait anxiety, and trait stress ($r = .40$, $p < .001$; $r = .40$, $p < .001$; $r = .23$, $p = .017$; respectively). To test this hypothesis more thoroughly and to test which predictor would remain robust when controlling for the other variables, we then conducted a regression analysis including all three dispositional variables as predictors for general state anxiety. *R* for this regression analysis was significantly different from zero, $F(3,104) = 8.64$, $R^2 = .199$, $R^2_{adj} = .176$, $p < .001$. Trait test anxiety was a significant predictor, $t(104) = 2.27$, $B = 0.30$, $SD = 0.13$, $\beta = .257$, $p = .025$. General trait anxiety was also significant, $t(104) = 2.35$, $B = 0.31$, $SD = 0.13$, $\beta = .270$, $p = .021$. In contrast, trait stress was not significant, $t(104) = -0.39$, $B = -0.05$, $SD = 0.12$, $\beta = -.042$, $p = .699$.

To test Hypothesis 6b, we then conducted another regression analysis using SPSS and PROCESS (model 1; Hayes, 2018) to predict general state anxiety using the learning condition, task-specific self-concept, general trait anxiety, as well as the interaction of the learning condition and general trait anxiety. We chose general trait anxiety because it was descriptively the strongest dispositional variable in the regression analysis presented above. *R* for this regression analysis was significantly different from zero, $F(4,103) = 10.02$, $R^2 = .280$, $R^2_{adj} = .252$, $p < .001$. The learning condition was a significant predictor, $t(103) = 3.12$, $B = 0.62$, $SD = 0.20$, $\beta = .271$, $p = .002$. The specific self-concept was also significant, $t(103) = -3.15$, $B = -0.32$, $SD = 0.10$, $\beta = -.276$, $p = .002$. General trait anxiety was also significant, $t(103) = 3.24$, $B = 0.46$, $SD = 0.14$, $\beta = .399$, $p = .002$. The interaction of the learning condition and general trait anxiety was, however, not significant, $t(103) = -0.48$, $B = -0.09$, $SD = 0.19$, $\beta = -.059$, $p = .630$.[7] These findings supported Hypotheses 4b and 5b but did not support Hypothesis 6b. Thus, the learning condition and the trait variables were linked to higher general state anxiety but there was no interaction of these situational and individual variables. Controlling for participants gender did not change any of the results of our second study.

3.3. Discussion

Our second study was conducted to explore the influence of tests compared to reading on participants' acute negative evaluations of such learning situations and on their acute stress experiences. Stress experiences were thereby indicated by self-reported state stress

[6] Using general trait anxiety or trait stress instead of trait test anxiety leads to mirroring findings. In each case, the learning condition and the task-specific self-concept were significant predictors, but neither the dispositional variable nor the interaction of the learning condition and the dispositional variable were significant.

[7] Using trait test anxiety or trait stress instead of general trait anxiety leads to mirroring findings. In each case, the learning condition, the task-specific self-concept, and the respective dispositional variable were significant predictors, but the interaction of the learning condition and the dispositional variable was not.

and general state anxiety (originally, participants pulse was also assessed as a physiological stress response but the analyses were—due to the issue with its interpretation described above—relocated to Appendix C). We further tested the effects of trait test anxiety, general trait anxiety, as well as trait stress and also tested interactions of the learning condition and these dispositional variables. Thus, the assumptions of Study 1 were investigated in a real learning situation in the laboratory.

The results showed that the learning condition significantly predicted—at least controlling for the task-specific self-concept—participants negative evaluation of the learning situation, state stress, and general state anxiety. As expected, tests constantly led to higher scores, thus, being evaluated as more negative and experienced as more stress- as well as anxiety-inducing than reading. Notably, these findings were obtained although we applied learning materials that were not completely new and not that difficult and even though the used single test was rather short, low-stake, and ungraded. Participants knew that their performance would neither change their payment nor influence their actual university courses. Nonetheless, retrieving previously learned information and solution steps to answer test questions with other participants nearby and shortly answering two questions aloud to receive feedback was pressuring enough to be negatively evaluated and to elicit more self-reported stress experiences compared to simply reading the same information. The results even remained robust when adding multiple dispositional variables and interaction-terms to the linear regression analyses. The interactions were never significant—probably due to the strong effects of the situation—and the dispositional variables were not continuously significant. For instance, none of the dispositional variables was significantly correlated to the negative evaluation of the learning situation.

There are, however, some limitations that need to be noted: E.g., the difference between the scale endpoints of the scale measuring state stress could have irritated participants. Further, the difference of the task-specific self-concept between participants in both learning conditions indicates that the randomization of participants was not ideal. Additionally—although we intentionally used mathematical tasks because a lot of university courses include mathematics, because mathematical tasks are cognitive demanding and realistic, and because beneficial effects of tests were previously found using such materials—the used learning materials could have been particularly stress- or anxiety-inducing. Mathematical tasks have, for instance, often been shown to be linked to anxiety and stress and may be especially threatening (e.g., Ashcraft & Moore, 2009). Thus, different learning materials may result in weaker effects, wherefore replications of the present work with different materials are valuable. Nonetheless, the descriptive statistics of both studies showed that state stress scores and the negative evaluation of the learning situation scores are extremely similar and comparable (the negative evaluation scores in Study 2 were descriptively even lower than the scores in Study 1). Hence, participants did not perceive and experience the learning situation and the test condition in Study 2 as more threatening than Study 1, indicating that the learning test in Study 2 still serves as a low-stake situation. Furthermore, one of the most important limitations is that we were unfortunately not able to test our hypotheses concerning participants pulse due to three—and potentially more—incorrectly self-reported pulse scores (see Appendix C for the analyses). Although the pulse measurement was—or should have been—rather easy and it is possible that no additional scores were wrong as well, we think that the results described in the Appendix must be considered with caution. Thus, it would be valuable for future work to test if the effects found in our second study can be transferred to physiological measures. Future studies could therefore use simpler or more detailed instructions for the pulse measurements, could additionally assess a baseline of participants pulse, or could directly apply different physiological metrics of stress experiences like cortisol, which may be even stronger and more robust than pulse measurements. Finally, we conducted only one session and only one rather short learning phase; thus, there is no indication of how long-lasting these negative effects of tests are and if they differ or evolve over time.

4. General discussion

The aim of our studies was to test linkages among tests as difficult learning tasks, dispositional variables, potential interactions of the learning tasks and dispositional variables, and learners' acute evaluations of and experiences in such learning situations. Thus, our studies are further steps to understand possible negative consequences of tests not as examinations but as low-stake learning tasks. Especially the explored—mostly non-significant—interactions are an important unique contribution of our work. Positively, we focused on participants' appraisals, attitudes, evaluations, and experiences in different learning situations (see Edwards & Templeton, 2005), thereby also taking multiple dispositional variables into account. Notably, these dispositional variables correlated differently to our dependent variables, supporting that the dependent variables assess different cognitive and affective responses due to acute stressful learning situations. Hence, another added value of our studies is that we assessed multiple dispositional variables as well as multiple dependent variables. Furthermore, the applied learning conditions—especially in Study 2, but also in Study 1—were rather realistic and operationalized to be similar to learning in actual university courses. We used, for instance, no monetary incentives to increase participants performance, worry, or motivation but relied on typical incentives in university settings like the existence of peers, potential external or social evaluations, or participants desire to do their best. We generally tried to conduct studies with materials and procedures that were as realistic as possible, so that implications of the findings could be easily transferred to actual classrooms and university courses. Most important, our research highlights that non-cognitive responses to learning tests and desirable difficulties—like evaluations, affects, and experiences—should also be considered in addition to later learning outcomes. We think that our work contributes to the existing literature and thereby surely stimulates future work and future important steps—for instance concerning the strongly related constructs of motivation, learning outcomes, and further triggered (undesirable) behaviour. Moreover, our studies certainly bridge literature focusing on (the effectiveness of) desirable difficulties and literature focusing on stress and anxiety (and its influences on later learning outcomes), hence, connecting these research fields.

Notably, our two studies focusing on normally beneficial learning tests as instantiations of desirable difficulties share a lot of similarities and connections: For instance, the operationalizations of both studies followed the typical pattern of tests as desirable difficulties and both tried to depict the specific learning situation as realistically as possible, among others by applying (or describing)

complex and curricular materials and situations, by including varying test questions, and through using only incentives that are actually present in university courses. Thus, both studies include some aspects of (evaluation, monitoring, or outcome) pressure that learners normally encounter in their courses (e.g., intrinsic motivation to do their best or to make a good impression on others as well as the actual or imagined existence of peers and lecturers). Nonetheless, all test situations in the two studies are low-stake (e.g., because all tests were ungraded, all learners worked simultaneously on the same tasks, there were no monetary incentives, the learned information in Study 1 were only hypothetical and not specified, and the learned information in Study 2 were not consequential for participants actual courses, …). Moreover, the non-significant results regarding stress experiences in Study 1—argued to be due because multiple (hypothetical) learning tests over a whole semester might be difficult to imagine—led to the usage of one distinct learning situation in Study 2 and to the measurement of stress experiences immediately after this single learning situation. Also following these non-significant effects on state stress in Study 1, we then included even more instruments assessing different indicators of stress experiences in Study 2, so that we were able to analyse even further aspects of immediate stress responses. Hence, the two studies building the present work and their respective methods share a lot of similarities but also some differences (e.g., concerning the number of learning conditions, the setting of the respective study, the country in which the samples were recruited, and the applied or described learning materials) but can, nonetheless, together contribute to the existing literature.

Study 1 found that imagined learning scenarios including tests were evaluated as more negative than imagined learning situations including re-reading. This remained robust even when adding dispositional variables. Additionally, trait test anxiety and trait stress were positively linked to participants negative evaluations of the learning scenarios and to participants stress experiences in such imagined situations. There was a significant interaction between trait stress and the dummy-coded variable tests private, insofar as that participants with average and high trait stress in the tests with private results learning scenario condition rated the learning situation as more negative than those participants in the re-reading control learning scenario condition. The negative evaluation of the learning situation did, however, not differ between participants with lower trait stress in these two scenario conditions. In Study 2, taking a learning test—in contrast to a reading task—in a laboratory setting with actual learning materials also led to a more negative evaluation of the learning situation and additionally even to more stress experiences indicated by more self-reported state stress and higher general state anxiety. These effects remained robust when adding dispositional variables like trait test anxiety, general trait anxiety, and trait stress. Additionally, some of these dispositional variables were also positively linked to participants' stress experiences; however, the effects of the dispositional variables were not as continuously and not as robust as the effect of the learning condition. There were no significant interactions.

Taken together, the results of both studies show that imagined and real learning tests negatively impact learners' cognitive evaluations as well as attitudes and that actual learning tests applying real learning materials also lead to more negative affective stress responses. Notably, dispositional trait variables in both studies were also linked to negative evaluations of the learning situation and stress experiences—however, effects of the trait variables were stronger for the hypothetical scenarios, whereas effects of the learning condition were stronger for actual learning situations. Interestingly, the effect of the learning condition on the dependent variables was continuously significant in Study 2 but not continuously significant in Study 1—nonetheless, the descriptive statistics of the state stress scores and the negative evaluation of the learning situation scores were extremely similar in both studies. Thus, effects of imagined scenarios seem to be comparable to effects of actual learning in a laboratory setting.

Apart from the one interaction found in Study 1, serving as evidence for potential interactions between situational and individual factors, we otherwise repeatedly found no support for interactions, indicating that difficult learning situations using tests are threatening for most individuals and not only for those already scoring higher on dispositional stress or anxiety. Thus, reporting such non-significant findings is valuable. It could be possible that the interaction in Study 1 was only found due to our operationalization—scenarios in general and imagined tests with private results specifically—and that there really are no interactions in laboratory and classroom settings. There were, for instance, also no interactions in the laboratory study from Hinze and Rapp (2014). Thus, more studies testing interactions are valuable.

Notably, the described negative consequences arose although the tests in Study 1 consisted only of hypothetical tests in imagined scenarios and the single test in Study 2 was ungraded, low-stake, and without consequences dependent of participants performance. Thus, even tests solely used as learning situations are acute stressors that can result in negative side-effects and even tests applied as low-stake can still be perceived as high-stake or pressuring. This is important for the practical application of learning tests in school or university contexts: Lecturers and learners alike must know that tests can lead to immediate negative evaluations and stress experiences for (almost) every learner and not only for those already high on dispositional stress and anxiety. Hence, lecturers must be careful when using tests because these can lead to negative consequences: On the one hand, increased stress experiences indicated by higher situational stress and anxiety as well as increased negative evaluations of the situation—including, for instance, more feelings of anger, uncertainty, and nervousness as well as more perceived injustice and unfairness—are negative and undesirable states that individuals normally try to avoid. On the other hand, such negative consequences can, in turn, additionally also lead to further negative side-effects: Previous work showed that stress, anxiety, pressure, perceptions of situations as threatening, and negative situation evaluations were linked to cognitive deficits, health problems, lower persistence, lower effort, lower motivation to learn, more exhaustion, and more negative affect (e.g., DeLongis et al., 1988; Leiner et al., 2018; LePine et al., 2004; Struthers et al., 2000). LePine et al. (2004), for instance, argued in line with the transactional theory of stress (e.g., Lazarus, 1990) that especially stress appraised as a hindrance (or a threat) is linked to negative situation perceptions, decreased motivation to learn, less effort focussed on the task, and withdrawal from the situation. Other work also showed that especially positive affective experiences were linked to more cognitive activity, expectancy of success, and achievement (see Buff, Reusser, Rakoczy, & Pauli, 2011)—thus, it seems important to focus on learners' experiences, affects, and evaluations. Hence, negative consequences due to tests could generally trigger further undesirable outcomes and future work could specifically test which consequences the found effects have, for instance, on learners' motivation and behaviour.

There are also limitations and future directions of both studies that we care to discuss. For instance, we argued in the beginning that the inconsistent previous findings are a result of the time at which the dependent variables are measured and that some of the studies did not measure acute stress but long-term effects. Notably, we only supposed but did not manipulate or test this assumption—thus, future work focusing thereon would be valuable. It would also be interesting to further explore the scale we used to measure participants negative evaluation of the learning situation: Although the used ten items—and the applied indicators of stress experiences—serve as a good overview of different aspects of evaluations and situation perceptions, the individual items could still differ among each other regarding their actual negativity or undesirability. Thus, future work could try to validate or optimize the applied measurement. Additionally, as mentioned in the introduction section, beneficial effects of tests arise when using a plethora of tests, test question formats, and learning materials, however, we were only able to apply two learning scenarios including slightly differing tests in Study 1 and only one actual learning test in Study 2. Thus, future work should operationalize tests in even more variations to explore if the found effects arise for all sorts of learning tests. Moreover, instead of using tests as the instantiation of desirable difficulties, future studies should check if the same results arise when using different difficulties like generation or disfluency. Furthermore, future studies should additionally examine if the observed negative consequences of tests in turn moderate the effectiveness of difficult learning situations because previous studies found anxiety and stress (among others due to reduced motivation to learn) to be negatively linked to academic achievement, learning outcomes, and performance (e.g., Chen & Chang, 2009; Hinze & Rapp, 2014; LePine et al., 2004; Mok & Chan, 2016; Seipp, 1991; Struthers et al., 2000; Weissgerber & Reinhard, 2018). However, some work also resulted in differing findings: For instance, LePine et al. (2004) showed that although stress associated with hindrances in the learning context was negatively related to motivation to learn and to learning performance, stress associated with challenge was in contrast positively related to motivation to learn and learning performance. Sung, Chao, and Tseng (2016) also found a positive link between test anxiety and performance, indicating that higher anxiety was generally linked to higher performance. Another study showed that the benefits of a test, compared to a control condition, decreased with higher test anxiety—but only for participants with lower working memory capacities and not for participants with higher working memory capacities (Tse & Pu, 2012). In contrast, Messineo, Gentile, and Allegra (2015) found that—although learners with high test anxiety generally performed less well than those with lower test anxiety—they benefited more from learning tests than learners with lower test anxiety. Hence, further work is still valuable. This also applies to further assessments of learners' motivation and to potential linkages among learning tests, stress, and learners' motivation to learn, to increase their effort, and to persist while working on such difficult learning tasks. Finally, future studies could also directly try to reduce negative consequences of tests; e.g., using mindfulness interventions, emotion regulation techniques, coping strategies, or through priming the effectiveness of tests (e.g., Jamieson et al., 2016; Khng, 2017; Struthers et al., 2000; Zenner, Herrnleben-Kurz, & Walach, 2014).

5. Conclusion

Summarizing we can say that—although difficult learning situations like tests are normally beneficial for learning—they can also lead to negative consequences. These include more negative evaluations of the learning situations and higher stress experiences indicated by state stress and general state anxiety. Dispositional variables like test anxiety, general trait anxiety, and trait stress were also positively linked to such negative consequences but not as robust as the learning situation. We found only one significant interaction, all other potential interactions between the learning condition and the assessed dispositional variables were continuously not significant. Hence, lecturers and learners alike must be aware that tests can generally trigger negative consequences.

Funding

This research was supported by a LOEWE grant from the Hessian Ministry for Science and the Arts entitled "desirable difficulties; intrinsic cognitive motivation and performance expectancies" awarded to Marc-André Reinhard.

CRediT authorship contribution statement

Kristin Wenzel: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing. **Marc-André Reinhard:** Funding acquisition, Conceptualization, Formal analysis, Validation, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

We thank Chawwah Grünberg, Sonja Haverland, Antonia Mariß, Nina Reinhardt, Luisa Neufeld, Celina Stolz, Tobias Steppat, and Laura Wagner for their help in recruiting and data collection.

Appendix A

Tests with public results learning scenario condition (including the instructions):

"This is a potential scenario that could happen in your daily life as a student. We would like to ask you to transport yourself in the situation, and to imagine it as strongly as you can. Imagine that you are a student in college and have lots of exams to write. During one of your majors your professor tries to increase you, and your fellow students learning success, and enhance your chance to pass the exam. Therefore, half an hour before the end of every session you write an ungraded test, and answer multiple questions concerning the content of that session. Once the half an hour is up you can go home. Shortly following every session all students receive an e-mail with the matriculation numbers of everyone, and their test results, ranking from best to worst."

Tests with private results learning scenario condition (including the instructions):

"This is a potential scenario that could happen in your daily life as a student. We would like to ask you to transport yourself in the situation, and to imagine it as strongly as you can. Imagine that you are a student in college and have lots of exams to write. During one of your majors your professor tries to increase you, and your fellow students learning success, and enhance your chance to pass the exam. Therefore, half an hour before the end of every session you write an ungraded test, and answer multiple questions concerning the content of that session. Once the half an hour is up you can go home. Shortly following every session you receive a private e-mail with your own test results."

Re-reading control learning scenario condition (including the instructions):

"This is a potential scenario that could happen in your daily life as a student. We would like to ask you to transport yourself in the situation, and to imagine it as strongly as you can. Imagine that you are a student in college and have lots of exams to write. During one of your majors your professor tries to increase you and your fellow students learning success and enhance your chance to pass the exam. Therefore, half an hour before the end of every session your professor hands you a summary with all the relevant information of that session. In this time you read the materials. Once the half an hour is up you can go home."

Negative Evaluation of the Learning Situation: Instructions and items

"Please answer the following questions according to your imagined mood/ perception/ thoughts/ feelings **during the situation displayed in the scenario**."

"Concerning the imagined scenario, …

1 How strenuous did you find the described and imagined learning-situation? One (*not strenuous at all*) – seven (*extremely strenuous*)
2 How (un)just did you find the described and imagined way of learning in the situation? One (*extremely unjust*) – seven (*extremely just*), *recoded item
3 How difficult would your rate the learning in the described situation? One (*not difficult at all*) – seven (*extremely difficult*)
4 How fair or unfair would you rate the way of learning in such a situation? One (*extremely unfair*) – seven (*extremely fair*), *recoded item
5 How angry would you be if you were in such a situation and had to learn in such a manner? One (*not in the least bit angry*) – seven (*extremely angry*)
6 How relaxing would you rate such a learning-situation? One (*not relaxing at all*) – seven (*extremely relaxing*), *recoded item
7 How overstrained would you feel if you were in such a learning-situation? One (*not at all*) – seven (*totally*)
8 How annoyed would you feel if you were in such a learning-situation? One (*not at all*) – seven (*extremely*)
9 How uncertain would you feel if you had to learn in a way as described in the situation? One (*not at all*) – seven (*extremely*)
10 How inappropriate would you rate such a learning-situation? One (*not inappropriate at all*) – seven (*extremely inappropriate*)

Appendix B

Example items of Study 2 (translated for this presentation, used materials in German)

Learning task in the reading control condition:

a) (0|0); (2|2) fits the [blue] line and has the functional equation [y = x] with the slope m=

[1]; the point (22|22) is ☒ is not ☐ on this line, because [22=1·22] or [1=1].

Correct solution:

Step 1: Insert the given points (2|2) and (0|0) into the formula to compute the slope.

$m = \frac{y}{x} = \frac{22}{}$ | Recap, if you enter (2|2) and (0|0) in the general formula to compute the slope $m = \frac{y_2 - y_1}{x_2 - x_1}$, than you will obtain $= \frac{2-0}{2-0} = \frac{2}{2} = \frac{y}{x}$

$m = 1$

Step 2: This computed slope of $m = 1$ is inserted into the equation defining the line through the origin of $y = m \cdot x$.

$y = 1 \cdot x$| Insert $m = 1$ into $y = m \cdot x$

$y = x$

Step 3: Enter the new unknown point (22|22) with a known point on the line, e. g. (0|0) or (2|2) into the formula to compute the slope.

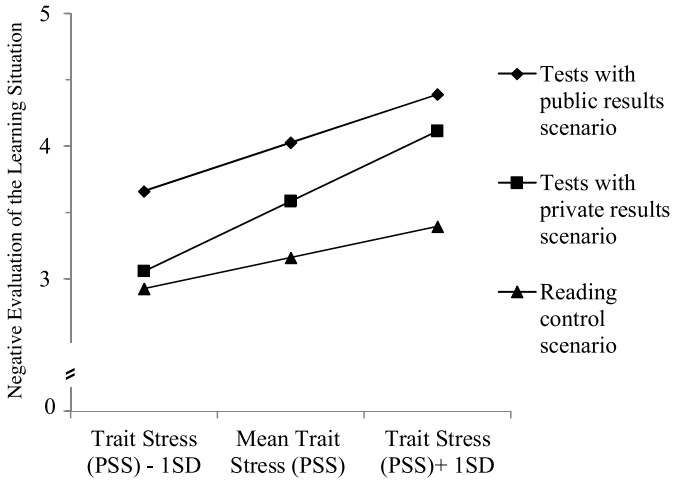


Fig. C1. The conditional effect of the learning scenario condition on the negative evaluation of the learning situation for participants with relatively low (−1SD), average, and relatively high (+1SD) trait stress measured with the PSS in Study 1.
Note. The learning scenario condition was dummy-coded: dummy variable 1: *tests private*, 1 = tests with private results learning scenario condition; dummy variable 2: *tests public*, 1 = tests with public results learning scenario condition; reference category: 0 = re-reading control learning scenario condition.

$m = \frac{22-022-0}{} = \frac{y}{x} = \frac{22}{22} = 1$ *or* $m = \frac{22-2}{22-2} = \frac{20}{20} = 1$

Step 4: Compare, if your computed slope based on an unknown and known point in step 3 matches the previously computed slope in step 1.

$m = 1$ stemming from the functional equation of $y = 1 \cdot x$ in step 1

$m = \frac{22}{22} = 1$ stemming from the computed slope of an unknown and known point in step 3

Both slopes of $m = \frac{22}{22} = 1$ and of $m = 1$ are equal, that is $1 = 1$. Therefore, the unknown point has to fit the blue line.

alternatively:

Enter the unknown point of (22|22) into your derived functional equation from step 2 and check, whether the result is a true mathematical statement.

$y = m \cdot x$ | Point (22|22) has $x = 22$ and $y = 22$

$22 = 1 \cdot 22$

$22 = 22$| This is true. Thus, the unknown point (22|22) has to fit the blue line.

Learning task in the test condition:

a) (0|0); (2|2) fits the ☐ line and has the functional equation ☐ with the slope m=

☐ point (22|22) ☐ is ☐ is not on this line, because ☐ or ☐.

Appendix C

Further Analyses

Study 1

Third figure depicting the significant interaction of tests private and trait stress (Fig. C1)

Study 2

The relegated information, analyses, and interpretations of the dependent variable pulse as an indicator of physiological stress experiences:

See the methods section of Study 2 concerning the issues why this dependent variable was relegated and why the following analyses have to be interpreted with caution.

Due to the three participants that did not (correctly) report their pulse (see the methods section of Study 2), the following analyses were conducted with a sample of $N = 105$ (test condition: $n = 51$; reading control condition: $n = 54$; descriptive statistics of pulse: $M = 80.56$, $SD = 14.65$, range: 51–141) and have to be interpreted with caution. To test Hypothesis 4c, we conducted a *t*-test to compare participants stress experiences indicated by their average pulse in both learning conditions: $M_{test} = 82.33$, $SD_{test} = 16.28$, $M_{reading} = 78.89$, $SD_{reading} = 12.86$, $t(103) = 1.21$, $p = .230$. There was no significant difference. In line with the other analyses in the results section of Study 2, we again conducted a regression analysis predicting pulse through the learning condition (0 = reading control condition, 1 = test condition) and the task-specific self-concept. *R* for this regression analysis was not significantly different

from zero, $F(2,102) = 0.75$, $R^2 = .014$, $R^2_{adj} = -.005$, $p = .477$. The learning condition was not a significant predictor, $t(102) = 1.13$, $B = 3.31$, $SD = 2.93$, $\beta = .114$, $p = .261$. The task-specific self-concept was also not significant, $t(102) = 0.22$, $B = 0.33$, $SD = 1.46$, $\beta = .022$, $p = .825$.

Regarding Hypothesis 5c, we firstly regarded the correlations of pulse and the three dispositional variables trait test anxiety, general trait anxiety, and trait stress ($r = .01$, $p = .925$; $r = -.08$, $p = .394$; $r = -.15$, $p = .122$; respectively). To test this hypothesis more thoroughly and to test if any of the predictors would become significant when controlling for the other variables, we then conducted a regression analysis including all three dispositional variables as predictors for participants' pulse. R for this regression analysis was not significantly different from zero, $F(3,101) = 1.27$, $R^2 = .036$, $R^2_{adj} = .008$, $p = .288$. Trait test anxiety was not a significant predictor, $t(101) = 1.18$, $B = 2.18$, $SD = 1.85$, $\beta = .148$, $p = .240$. General trait anxiety was also not significant, $t(101) = -0.32$, $B = -0.65$, $SD = 1.99$, $\beta = -.043$, $p = .746$. Trait stress was also not significant, $t(101) = -1.60$, $B = -3.07$, $SD = 1.91$, $\beta = -.207$, $p = .112$.

To test Hypothesis 6c, we conducted another regression analysis predicting pulse using the learning condition, task-specific self-concept, trait test anxiety, as well as the interaction of the learning condition and trait test anxiety. We chose trait test anxiety in line with the other analyses in Study 2. R for this regression analysis was not significantly different from zero, $F(4,100) = 0.74$, $R^2 = .029$, $R^2_{adj} = -.010$, $p = .566$. None of the predictors was significant: Neither the learning condition, $t(100) = 1.11$, $B = 3.27$, $SD = 2.95$, $\beta = .112$, $p = .269$, nor the task-specific self-concept, $t(100) = 0.12$, $B = 0.17$, $SD = 1.50$, $\beta = .012$, $p = .908$, nor trait test anxiety, $t(100) = -0.72$, $B = -1.37$, $SD = 1.92$, $\beta = -.093$, $p = .476$. The interaction of the learning condition and trait test anxiety was also not significant, $t(100) = 1.22$, $B = 3.62$, $SD = 2.98$, $\beta = .156$, $p = .227$.[8] These findings did neither support Hypothesis 4c, Hypothesis 5c, nor Hypothesis 6c because none of our variables was linked to participants' pulse.

In contrast to our hypotheses, neither the learning condition, nor the dispositional variables, nor the interaction-terms were linked to participants' pulse. Seemingly, the pulse assessment—although being described in detail and with an experimenter present that participants could ask for help if they were not able to measure their pulse—was not as easy as we supposed. It is possible that other scores (additionally to the three impossible scores 0, 2.5, and 4 mentioned in the methods section of Study 2) were wrong as well but were not detected, thus, contorting the results and explaining why there were no significant effects. However, it is of course also possible that there simply are no effects of the learning condition or the trait variables on pulse. Thus, a low-stake learning situation including a test—although being evaluated as more negative and experienced as more stressful—might not be threatening enough to increase learners' pulse. Such learning tests might only be influential for subjective self-reports, cognitive appraisals, as well as affective states. Future studies should nonetheless use simpler or even more detailed instructions when assessing participants pulse. Moreover, measuring a baseline of participants pulse—assessed prior to learning—would also have been helpful for conducting a more valid indicator for changes in participants' heart rates—although we were only interested in differences in pulse between participants in both learning conditions and not in individual changes. In line with this, future work could also use different physiological metrics of stress like cortisol, which may be stronger and more robust.

References

Abouserie, R. (1994). Sources and levels of stress in relation to locus of control and self esteem in university students. *Educational Psychology, 14*(3), 323–330. https://doi.org/10.1080/0144341940140306.

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659–701. https://doi.org/10.3102/0034654316689306.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology General, 136*(4), 569–576. https://doi.org/10.1037/0096-3445.136.4.569.

Ashcraft, M. H., & Moore, A. M. (2009). Mathematics anxiety and the affective drop in performance. *Journal of Psychoeducational Assessment, 27*(3), 197–205. https://doi.org/10.1177/0734282908330580.

Beilock, S. L. (2008). Math performance in stressful situations. *Current Directions in Psychological Science, 17*(5), 339–343. https://doi.org/10.1111/j.1467-8721.2008.00602.x.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*(2), 201–210. https://doi.org/10.3758/BF03193441.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. Shimamura (Eds.), *Metacognition: knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher, & A. Koriat (Eds.), *Attention and performance. Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). The MIT Press.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society, 2*, 59–68.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From learning processes to cognitive processes: Essays in honor of William K. Estes, 2* (pp. 35–67).

Bradley, R. T., McCraty, R., Atkinson, M., Tomasino, D., Daugherty, A., & Arguelles, L. (2010). Emotion self-regulation, psychophysiological coherence, and test anxiety: results from an experiment using electrophysiological measures. *Applied Psychophysiology and Biofeedback, 35*(4), 261–283. https://doi.org/10.1007/s10484-010-9134-x.

Brougham, R. R., Zail, C. M., Mendoza, C. M., & Miller, J. R. (2009). Stress, sex differences, and coping strategies among college students. *Current Psychology, 28*(2), 85–97. https://doi.org/10.1007/s12144-009-9047-0.

[8] Using general trait anxiety or trait stress instead of trait test anxiety leads to mirroring findings. In each case none of the predictors was significant.

Buff, A., Reusser, K., Rakoczy, K., & Pauli, C. (2011). Activating positive affective experiences in the classroom: "Nice to have" or something more? *Learning and Instruction, 21*(3), 452–466. https://doi.org/10.1016/j.learninstruc.2010.07.008.

Bystritsky, A., & Kronemyer, D. (2014). Stress and anxiety: Counterpart elements of the stress/anxiety complex. *Psychiatric Clinics, 37*(4), 489–518. https://doi.org/10.1016/j.psc.2014.08.002.

Cassady, J. C. (2004a). The influence of cognitive test anxiety across the learning–testing cycle. *Learning and Instruction, 14*(6), 569–592. https://doi.org/10.1016/j.learninstruc.2004.09.002.

Cassady, J. C. (2004b). The impact of cognitive test anxiety on text comprehension and recall in the absence of external evaluative pressure. *Applied Cognitive Psychology, 18*(3), 311–325. https://doi.org/10.1002/acp.968.

Chen, I., & Chang, C. C. (2009). Cognitive load theory: An empirical study of anxiety and task performance in language learning. *Electronic Journal of Research in Educational Psychology, 7*(2), 729–746.

Clark, D., & Linn, M. C. (2003). Designing for knowledge integration: The impact of instructional time. *Journal of the Learning Sciences, 12*(4), 451–493. https://doi.org/10.1207/S15327809JLS1204_1.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24*(4), 385–396. https://doi.org/10.2307/2136404.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438–481. https://doi.org/10.3102/00346543058004438.

DeLongis, A., Folkman, S., & Lazarus, R. S. (1988). The impact of daily stress on health and mood: Psychological and social resources as mediators. *Journal of Personality and Social Psychology, 54*(3), 486–495. https://doi.org/10.1037//0022-3514.54.3.48.

Dickhäuser, O., & Reinhard, M. A. (2006). Daumenregel oder Kopfzerbrechen? [Rule of thumb or contemplating?]. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie, 38*(2), 62–68. https://doi.org/10.1026/0049-8637.38.2.62.

Dickhäuser, O., Schöne, C., Spinath, B., & Stiensmeier-Pelster, J. (2002). Die Skalen zum akademischen Selbstkonzept: Konstruktion und Überprüfung eines neuen Instrumentes [The academic self-concept scales: Construction and evaluation of a new instrument]. *Zeitschrift für differentielle und diagnostische Psychologie: ZDDP, 23*(4), 393–405. https://doi.org/10.1024//0170-1789.23.4.393.

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the: Effects of disfluency on educational outcomes. *Cognition, 118*(1), 111–115. https://doi.org/10.1016/j.cognition.2010.09.012.

Dobson, J. L., & Linderholm, T. (2015). The effect of selected "desirable difficulties" on the ability to recall anatomy information. *Anatomical Sciences Education, 8*(5), 395–403. https://doi.org/10.1002/ase.1489.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266.

Edwards, J. A., & Templeton, A. (2005). The structure of perceived qualities of situations. *European Journal of Social Psychology, 35*(6), 705–723. https://doi.org/10.1002/ejsp.271.

Endler, N. S. (1997). Stress, Anxiety and coping: The multidimensional interaction model. *Canadian Psychology/Psychologie Canadienne, 38*(3), 136–153. https://doi.org/10.1037/0708-5591.38.3.136.

Epel, E. S., Crosswell, A. D., Mayer, S. E., Prather, A. A., Slavich, G. M., Puterman, E., et al. (2018). More than a feeling: A unified view of stress measurement for population science. *Frontiers in Neuroendocrinology, 49*, 146–169. https://doi.org/10.1016/j.yfrne.2018.03.001.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149.

Fazio, L. K. (2017). The effects of retrieval practice on fraction arithmetic knowledge. *Cognitive development from a strategy perspective: A festschrift for Robert Siegler* (pp. 169–182).

Fliege, H., Rose, M., Arck, P., Levenstein, S., & Klapp, B. F. (2001). Validierung des "perceived stress questionnaire" (PSQ) an einer deutschen Stichprobe. [Validation of the "Perceived Stress Questionnaire"(PSQ) in a German sample]. *Diagnostica, 47*(3), 142–152. https://doi.org/10.1026//0012-1924.47.3.142.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis* (second edition). The Guilford Press.

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: Performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology, 28*(4), 597–606. https://doi.org/10.1002/acp.3032.

Hobfoll, S. E. (1989). Conservation of resources: A new attempt at conceptualizing stress. *The American Psychologist, 44*(3), 513–524. https://doi.org/10.1037/0003-066X.44.3.513.

Hodapp, V., Rohrmann, S., & Ringeisen, T. (2011). *Prüfungsangstfragebogen.* Hogrefe: PAF.

Hoferichter, F., Raufelder, D., Ringeisen, T., Rohrmann, S., & Bukowski, W. M. (2016). Assessing the multi-faceted nature of test anxiety among secondary school students: An English version of the German test anxiety questionnaire: PAF-E. *The Journal of Psychology, 150*(4), 450–468. https://doi.org/10.1080/00223980.2015.1087374.

Jamieson, J. P., Peters, B. J., Greenwood, E. J., & Altose, A. J. (2016). Reappraising stress arousal improves performance and reduces evaluation anxiety in classroom exam situations. *Social Psychological and Personality Science, 7*(6), 579–587. https://doi.org/10.1177/1948550616644656.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review, 27*(2), 317–326. https://doi.org/10.1007/s10648-015-9309-3.

Karpicke, J. D., & Roediger, H. L., III (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology Learning, Memory, and Cognition, 33*(4), 704–719. https://doi.org/10.1037/0278-7393.33.4.704.

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own? *Memory, 17*(4), 471–479. https://doi.org/10.1080/09658210802647009.

Kausar, R. (2010). Perceived stress, academic workloads and use of coping strategies by university students. *Journal of Behavioural Sciences, 20*(1), 31–45.

Khanna, M. M. (2015). Ungraded pop quizzes: Test-enhanced learning without all the anxiety. *Teaching of Psychology, 42*(2), 174–178. https://doi.org/10.1177/0098628315573144.

Khanna, M. M., & Cortese, M. J. (2016). The benefits of quizzing in content-focused versus skills-focused courses. *Scholarship of Teaching and Learning in Psychology, 2*(1), 87–97. https://doi.org/10.1037/stl0000051.

Khng, K. H. (2017). A better state-of-mind: Deep breathing reduces state anxiety and enhances test performance through regulating test cognitions in children. *Cognition & Emotion, 31*(7), 1502–1510. https://doi.org/10.1080/02699931.2016.1233095.

Laux, L., Glanzmann, P., Schaffner, P., & Spielberger, C. D. (1981). *Das state-trait-angstinventar: STAI [The state-trait anxiety inventory: STAI]*. Beltz.

Lay, C. H., Edwards, J. M., Parker, J. D., & Endler, N. S. (1989). An assessment of appraisal, anxiety, coping, and procrastination during an examination period. *European Journal of Personality, 3*(3), 195–208. https://doi.org/10.1002/per.2410030305.

Lazarus, R. S. (1990). Theory-based stress measurement. *Psychological Inquiry, 1*(1), 3–13. https://doi.org/10.1207/s15327965pli0101_1.

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping.* Springer Publishing Company.

Lazarus, R. S., & Folkman, S. (1987). Transactional theory and research on emotions and coping. *European Journal of Personality, 1*(3), 141–169. https://doi.org/10.1002/per.2410010304.

Leiner, J. E. M., Scherndl, T., & Ortner, T. M. (2018). How do men and women perceive a high-stakes test situation? *Frontiers in Psychology, 9*, 2216. https://doi.org/10.3389/fpsyg.2018.02216.

LePine, J. A., LePine, M. A., & Jackson, C. L. (2004). Challenge and hindrance stress: Relationships with exhaustion, motivation to learn, and learning performance. *The Journal of Applied Psychology, 89*(5), 883–891. https://doi.org/10.1037/0021-9010.89.5.883.

Levenstein, S., Prantera, C., Varvo, V., Scribano, M. L., Berto, E., Luzi, C., et al. (1993). Development of the Perceived Stress Questionnaire: A new tool for psychosomatic research. *Journal of Psychosomatic Research, 37*(1), 19–32. https://doi.org/10.1016/0022-3999(93)90120-5.

Lim, S. W. H., Ng, G. J. P., & Wong, G. Q. H. (2015). Learning psychological research and statistical concepts using retrieval-based practice. *Frontiers in Psychology, 6*, 1484. https://doi.org/10.3389/fpsyg.2015.01484.
Lumley, M. A., & Provenzano, K. M. (2003). Stress management through written emotional disclosure improves academic performance among college students with physical symptoms. *Journal of Educational Psychology, 95*(3), 641–649. https://doi.org/10.1037/0022-0663.95.3.641.
Lundqvist, J. (2019). *Individual differences in working memory and the effect of retrieval practice* (Master thesis, Umea University).
Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology, 38*(2), 94–97. https://doi.org/10.1177/0098628311401587.
Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. (2019). How the amount and spacing of retrieval practice affect the short-and long-term retention of mathematics knowledge. *Educational Psychology Review*, 1–19. https://doi.org/10.1007/s10648-019-09489-x.
Maass, J. K., & Pavlik, P. I., Jr (2016). Modeling the influence of format and depth during effortful retrieval practice. *International Educational Data Mining Society*.
McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103*(2), 399. https://doi.org/10.1037/a0021782.
McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*(2), 200–206. https://doi.org/10.3758/BF03194052.
McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27*(3), 360–372. https://doi.org/10.1002/acp.2914.
McGrath, J. E. (1970). *Social and psychological factors in stress*. Holt, Rinehart, & Winston.
Messineo, L., Gentile, M., & Allegra, M. (2015). Test-enhanced learning: Analysis of an experience with undergraduate nursing students. *BMC Medical Education, 15*, 182. https://doi.org/10.1186/s12909-015-0464-5.
Mills, N., Pajares, F., & Herron, C. (2006). A reevaluation of the role of anxiety: Self-efficacy, anxiety, and their relation to reading and listening proficiency. *Foreign Language Annals, 39*(2), 276–295. https://doi.org/10.1111/j.1944-9720.2006.tb02266.x.
Mok, W. S. Y., & Chan, W. W. L. (2016). How do tests and summary writing tasks enhance long-term retention of students with different levels of test anxiety? *Instructional Science, 44*(6), 567–581. https://doi.org/10.1007/s11251-016-9393-x.
Nyroos, M., Schéle, I., & Wiklund-Hörnqvist, C. (2016). Implementing test enhanced learning: Swedish teacher students' perception of quizzing. *International Journal of Higher Education, 5*(4), 1–12. https://doi.org/10.5430/ijhe.v5n4p1.
O'Neil, J. H., Spielberger, C. D., & Hansen, D. N. (1969). Effects of state anxiety and task difficulty on computer-assisted learning. *Journal of Educational Psychology, 60*(5), 343–350. https://doi.org/10.1037/h0028323.
Pieger, E., Mengelkamp, C., & Bannert, M. (2016). Metacognitive judgments and disfluency–does disfluency lead to more accurate judgments, better control, and better performance? *Learning and Instruction, 44*(1), 31–40. https://doi.org/10.1016/j.learninstruc.2016.01.012.
Rauthmann, J. F. (2012). You say the party is dull, I say it is lively: A componential approach to how situations are perceived to disentangle perceiver, situation, and perceiver× situation variance. *Social Psychological and Personality Science, 3*(5), 519–528. https://doi.org/10.1177/1948550611427609.
Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., et al. (2014). The Situational Eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology, 107*(4), 677–718. https://doi.org/10.1037/a0037250.
Reeder, L. G., Schrama, P. G. M., & Dirken, J. M. (1973). Stress and cardiovascular health: An international cooperative study—I. *Social Science & Medicine (1967), 7*(8), 573–584. https://doi.org/10.1016/0037-7856(73)90026-7.
Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x.
Roelle, J., & Berthold, K. (2017). Effects of incorporating retrieval into learning tasks: The complexity of the tasks matters. *Learning and Instruction, 49*(1), 142–156. https://doi.org/10.1016/j.learninstruc.2017.01.008.
Rowland, C. A. (2014, August 25). The Effect of Testing Versus Restudy on Retention: A Meta-Analytic Review of the Testing Effect. *Psychological Bulletin*. Advance online publication. https:/doi.org/10.1037/a0037559.
Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology, 46*(4), 929. https://doi.org/10.1037/0022-3514.46.4.929.
Sarason, I. G., & Sarason, B. R. (1990). Test anxiety. *Handbook of social and evaluation anxiety* (pp. 475–495). Boston, MA: Springer. https://doi.org/10.1007/978-1-4899-2504-6_16.
Schulz, P., Schlotz, W., & Becker, P. (2004). *Trierer Inventar zum chronischen Stress: TICS [Trier inventory for chronic stress: TICS]*. Hogrefe.
Schunk, H. D., & Gaa, J. P. (1981). Goal-setting influence on learning and self-evaluation. *The Journal of Classroom Interaction, 16*(2), 38–44.
Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research, 4*(1), 27–41. https://doi.org/10.1080/08917779108248762.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). *A 21 word solution*. Available at SSRN: https://ssrn.com/abstract=2160588 or. https://doi.org/10.2139/ssrn.2160588.
Stiggins, R. J. (2001). The unfulfilled promise of classroom assessment. *Educational Measurement Issues and Practice, 20*(3), 5–15. https://doi.org/10.1111/j.1745-3992.2001.tb00065.x.
Struthers, C. W., Perry, R. P., & Menec, V. H. (2000). An examination of the relationship among academic stress, coping, motivation, and performance in college. *Research in Higher Education, 41*(5), 581–592. https://doi.org/10.1023/A:1007094931292.
Sung, Y. T., Chao, T. Y., & Tseng, F. L. (2016). Reexamining the relationship between test anxiety and learning achievement: An individual-differences perspective. *Contemporary Educational Psychology, 46*, 241–252. https://doi.org/10.1016/j.cedpsych.2016.07.001.
Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction, 12*(3), 185–233. https://doi.org/10.1207/s1532690xci1203_1.
Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States of America, 110*(16), 6313–6317. https://doi.org/10.1073/pnas.1221764110.
Tse, C. S., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *Journal of Experimental Psychology Applied, 18*(3), 253. https://doi.org/10.1037/a0029190.
Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review, 27*(2), 247–264. https://doi.org/10.1007/s10648-015-9310-x.
Weissgerber, S. C., & Reinhard, M.-A. (2018). Pilot study on the relationship of test anxiety to utilizing self-testing in self-regulated learning. *International Journal of Psychological Studies*. https://doi.org/10.5539/ijps.v10n4p95. Advance online publication.
Wenzel, K., & Reinhard, M.-A. (2019). Relatively unintelligent individuals do not benefit from intentionally hindered learning: The role of desirable difficulties. *Intelligence, 77*, Article 101405. https://doi.org/10.1016/j.intell.2019.101405.
Wenzel, K., & Reinhard, M.-A. (2020). Tests and academic cheating: Do learning tasks influence cheating by way of negative evaluations? *Social Psychology of Education, 23*, 721–753. https://doi.org/10.1007/s11218-020-09556-0.
Wirebring, L. K., Lithner, J., Jonsson, B., Liljekvist, Y., Norqvist, M., & Nyberg, L. (2015). Learning mathematics without a suggested solution method: Durable effects on performance and brain activity. *Trends in Neuroscience and Education, 4*(1–2), 6–14. https://doi.org/10.1016/j.tine.2015.03.002.
Wong, S. S. H., Ng, G. J. P., Tempel, T., & Lim, S. W. H. (2019). Retrieval practice enhances analogical problem solving. *Journal of Experimental Education, 87*(1), 128–138. https://doi.org/10.1080/00220973.2017.1409185.
Zenner, C., Herrnleben-Kurz, S., & Walach, H. (2014). Mindfulness-based interventions in schools—A systematic review and meta-analysis. *Frontiers in Psychology, 5*, 603. https://doi.org/10.3389/fpsyg.2014.00603.

APPENDIX E

Wenzel, K., & Reinhard, M.-A. (2020). Tests and academic cheating: do learning tasks influence cheating by way of negative evaluations?. *Social Psychology of Education, 23*(3), 721–753. https://doi.org/10.1007/s11218-020-09556-0

This is the final article version published by Springer Nature in *Social Psychology of Education* available online: https://link.springer.com/article/10.1007/s11218-020-09556-0

Social Psychology of Education (2020) 23:721–753
https://doi.org/10.1007/s11218-020-09556-0

Tests and academic cheating: do learning tasks influence cheating by way of negative evaluations?

Kristin Wenzel[1] · **Marc-André Reinhard**[1]

Received: 22 August 2019 / Accepted: 9 April 2020 / Published online: 27 April 2020

Abstract

Desirable difficulties like tests were often shown to increase long-term learning. However, due to the complexity and difficulty of such tasks, they are also argued to result in negative consequences like stress, anxiety, pressure, frustration, or negative evaluations. In other studies, such consequences were, in turn, often found to increase dishonest behaviour. Hence, the present work tests the assumptions that tests as difficult learning tasks, contrary to reading, lead to more negative evaluations of the learning situations, to more stress, and—directly and indirectly—to higher self-reported likelihoods of hypothetical cheating and to higher justifications for cheating. Thus, the learning situation itself, as well as negative consequences caused by the learning situation, is supposed to be linked to cheating. We conducted an online study in which participants read and imagined one of three hypothetical learning scenarios, either regarding one of two learning tests or a reading control task. Participants then rated negative consequences due to these scenarios, as well as likelihoods of cheating, and justifications for it, in a hypothetical examination. Our results showed no direct effects of the learning scenarios on likelihoods of hypothetical cheating or justifications. However, test scenarios were evaluated more negatively than the reading control scenario and these higher negative evaluations were in turn linked to higher likelihoods of own hypothetical cheating and to higher justifications. These findings indicate that tests as difficult learning tasks can indirectly influence cheating, at least in hypothetical scenarios. Future work should try to replicate and expand these results.

Keywords Tests · Cheating · Academic dishonesty · Desirable difficulties · Negative evaluations · Stress perceptions

✉ Kristin Wenzel
kristin.wenzel@uni-kassel.de

Marc-André Reinhard
reinhard@psychologie.uni-kassel.de

[1] Department of Psychology, University of Kassel, Holländische Straße 36-38, 34127 Kassel, Germany

1 Introduction

Challenging, difficult, and intentionally hindered learning tasks have often been shown to increase long-term learning outcomes compared to learning and processing that is fluent, easy, and simple, even though learners and lecturers normally assume the contrary (e.g., Bjork 1994; Bjork and Bjork 1992, 2011; Diemand-Yauman et al. 2011; Dobson and Linderholm 2015; Karpicke et al. 2009; Kornell et al. 2011). Previous work describes multiple incantations of such difficult learning tasks, for instance, *generation* (e.g., Bertsch et al. 2007), *distributed practice* (e.g., Cepeda et al. 2006), or *disfluency* (e.g., Diemand-Yauman et al. 2011). One of the most common desirable difficulties—and part of the present work—is, however, the application of learning tests or quizzes (also often called *testing effect*, *testing, learning/practice tests, test enhanced learning,* or *retrieval practice*): Taking a learning test on studied materials, after an initial study opportunity but before the final test or examination, increases retrieval of the learned information and enhances durable long-term learning as opposed to passively consumed and read materials (e.g., Adesope et al. 2017; Dobson and Linderholm 2015; McDaniel et al. 2007; Roediger and Karpicke 2006; Rowland 2014). These beneficial effects of tests were found in different settings (e.g., in laboratory, school, or university settings), for different ages (e.g., elementary school students, high school students, or university students), when using a broad array of materials or information (e.g., longer scientific textbook paragraphs, factual information, or lists of word-pairs like vocabulary), and when applying varying test question formats (e.g., multiple choice or short-answer questions inducing free recall, cued recall, or recognition; e.g., Adesope et al. 2017; Dobson and Linderholm 2015; Dunlosky et al. 2013; Roediger and Karpicke 2006; Rowland 2014).

Theoretically, the beneficial effects of desirable difficulties are attributed to stimulation of more elaborate cognitive processing, deeper semantic encoding, allocation of more resources, increased retention and transfer, strengthening of memory traces and associations, and anchoring of the learned information in long-term memory (e.g., Bjork 1994; Bjork and Bjork1992, 2011; Dunlosky et al. 2013; McDaniel et al. 1988; McNamara et al. 1996; Roediger and Karpicke 2006; Rowland 2014). Higher applied (cognitive) effort during retrieval and processing, increased quality and depth of processing and encoding (induced by retrieval attempts), higher amounts of cognitive capacities and resources utilized during information processing and retrieval, higher effort needed to solve the tasks, as well as generally higher difficulty and effort induced by both the test and the underlying retrieval practice are especially valuable for the positive effects of desirable difficulties (e.g., Bertsch et al. 2007; Bjork and Bjork 1992; Karpicke and Roediger 2007; Rowland 2014; Tyler et al. 1979). Difficult tasks also reduce learners' existing overconfidence and their *illusion of competence,* which otherwise convey the mistaken assumption that read information is already internalized: The learning test and the—due to the test—reduced competence illusion also enhance meta-cognitive accuracy of the hitherto learning process, in turn triggering the allocation of more resources and deeper, more elaborate, and more

systematic processing (e.g., Alter et al. 2007; Bjork 1999; Mihalca et al. 2017; Pieger et al. 2016).

Nonetheless, although desirable difficulties are argued to be beneficial, they are also by definition demanding, complicated, and challenging. Thus, lecturers in particular often express concern about the applicability and effectiveness of such intentionally hindered learning tasks for every individual (e.g., Diemand-Yauman et al. 2011; Lipowsky et al. 2015). In line with these concerns of lecturers, researchers also proposed that desirable difficulties are only beneficial for those individuals who can handle the needed increased effort, extended thought, and more elaborated and deeper processing, and for those who can correctly retrieve information and overcome the posed challenge (e.g., Alter et al. 2013; Kaiser et al. 2018; Kornell et al. 2011; Oppenheimer and Alter 2014; Richland et al. 2005; Rowland 2014). This, however, may not prove possible for everyone: Previous studies, for instance, showed that special requirements like higher previous knowledge, higher working memory capacity, higher intelligence, and higher reading ability are relevant skills for desirable difficulties to actually increase learning outcomes (e.g., Lehmann et al. 2016; McDaniel et al. 2002; McNamara et al. 1996; Wenzel and Reinhard 2019a). Hence, it is argued that desirable difficulties are not beneficial for every learner.

Notably, apart from that, we assume that difficult tasks used in learning contexts could also result in further negative side-effects: For instance, difficult learning tasks can sometimes pose too much additional demand (regarding, for instance, cognitive capacities, processing capacities, cognitive effort, or working memory capacities) as well as too much cognitive load on the learner, especially concerning authentic and more complex tasks and high element interactivity information (this applies in particular to learners with less expertise; see e.g., Kalyuga et al. 2001; Roelle and Berthold 2017; Sweller and Chandler 1994; van Gog and Sweller 2015; Wenzel and Reinhard 2019a). Because desirable difficulties are hard to solve and challenge learners' competence illusion and overconfidence (e.g., Bjork 1999), they are also assumed to reduce self-efficiency and to increase negative emotions, pressure, or fear of failure: Empirically, difficult tasks in general trigger perceptions of threat or anxiety and experiencing difficulties or giving incorrect answers feeds negatively into learners' self-perceptions (e.g., O'Neil et al. 1969; Sarason and Sarason 1990; Schunk and Gaa 1981). Besides, performing poorly—which can happen while working on desirable difficulties—leads to experiencing stress (e.g., Sarason and Sarason 1990; Schunk and Gaa 1981). Students also perceived difficult learning tasks and tasks that required more time and effort—and thus more workload—as more stress-inducing than easier tasks (e.g., Kausar 2010). Moreover, a laboratory study showed that learning tests resulted in more experienced pressure compared to a re-reading control task, even controlling for participants' dispositional anxiety (Hinze and Rapp 2014). Tests with high-stakes—induced by stating that monetary rewards were dependent of individuals' test results—were perceived as even more pressuring than tests with low-stakes in which monetary rewards were independent of individuals' test results. Thus, individuals felt some pressure simply from taking learning tests. Additionally, high-stake learning tests led to more anxiety than did low-stake tests and also negatively influenced participants' attitudes and interests (Hinze and Rapp 2014). Fittingly, participants in a laboratory setting that learned with a test,

compared to students that learned through reading the same information, evaluated the learning situation as more negative and experienced more stress and anxiety, even controlling for individual differences like trait stress and trait anxiety (Study 2, Wenzel and Reinhard 2019b). Moreover, due to the increased effort and the challenge learners must overcome when working with desirable difficulties in their courses, they are also argued to feel treated unfairly by their lecturers, in particular because they normally believe easy and fluent learning to be more effective (e.g., Karpicke et al. 2009; Kornell et al. 2011). Hence, we assume that difficult learning tasks should feel especially hard, pointless, and unfair.

Most important, these just described negative consequences (like stress, anxiety, or feelings of unfairness) were in other literature often assumed to be related to deceptive behaviour like academic cheating (e.g., Agnew 1992; Houser et al. 2012; Wowra 2007; see also the following paragraphs). Hence, applying desirable difficulties as learning tasks in universities could, directly and indirectly, cause more academic cheating and increase justifications for such cheating.

1.1 Academic cheating

People generally value honesty, trustworthiness, and credibility (e.g., Geißler et al. 2013), which is why they often refuse to admit their own cheating—or at least underreport it. Nonetheless, cheating behaviour can be observed throughout our daily lives (e.g., DePaulo et al. 1996; Feldman et al. 2002) and specifically in academic contexts (e.g., Finn and Frone 2004; McCabe 2001; McCabe et al. 2001; Simha and Cullen 2012; Whitley 1998). In one American survey, for instance, 74% of the participating students reported having seriously cheated on at least one test, while over 30% admitted repetitive and serious cheating in tests and exams (McCabe 2001; see also: Simha and Cullen 2012; Whitley 1998; Wowra 2007). However, actual numbers of academic cheating may be even higher because previous studies found imbalances between what students reported and what teachers actually observed in terms of cheating behaviour (e.g., Naghdipour and Emeagwali 2013). Typical incantations of such cheating behaviour in academic contexts include using cheat sheets in exams, copying answers in tests, relying on inappropriate collaboration during exams, or plagiarism (e.g., Jensen et al. 2002; Simha and Cullen 2012; Whitley 1998).

In general, different theories regarding cheating and deception do exist, very common theories being economic models like the *rational choice theory* (e.g., Akers 1990; Becker 1968) or the *strain theory* (e.g., Agnew 1992; Agnew and White 1992; Carmichael and Piquero 2004).

1.1.1 The rational choice theory

The rational choice theory describes the assumption that individuals decide whether or not to cheat after assessing possible gains or costs of such behaviour. Hence, the expected utility due to a cost–benefit calculation is important (e.g., Becker 1968). Dishonesty is mostly shown when the (for instance) financial or social gains of cheating

outweigh the costs of such behaviours, such as feelings of guilt and immorality or facing the consequences of getting caught. Potential gains of cheating in tests would include getting better grades, achieving better results with less effort, or making a good impression on others. However, people are also motivated to maintain a positive self-concept depicting themselves as moral, trustworthy, and honest (e.g., Abeler et al. 2019; Fischbacher and Föllmi-Heusi 2013; Mazar et al. 2008; Shalvi et al. 2011). Thus, individuals show higher degrees of dishonest behaviour when they feel entitled, deserving, or justified to do so (e.g., Cameron et al. 2008; Campbell et al. 2004; Fida et al. 2018; Mazar et al. 2008; Shalvi et al. 2011, 2015). Individuals can feel justified or entitled to behave dishonestly when, for instance, they can excuse deviant behaviour through denying their own responsibility (e.g., by blaming external forces like excessive workload), through criticising those who are at the receiving end of their dishonesty (e.g., by blaming them as unfair or unethical), or through rationalizing/normalizing their cheating behaviour (e.g., by stating that everybody cheats; see e.g., Olafson et al. 2013).

1.1.2 The strain theory

The strain theory further assumes that criminal or dishonest behaviour is influenced by negative affective states that result from perceived strain, strainful experiences, or stressors. Strain thereby includes failing to achieve, or being denied achieving, positive outcomes (like good grades); expecting or actually experiencing negative stimuli; perceiving a disjunction between aspirations or expectations and actual achievements/rewards; and experiencing a disjunction between fair or just outcomes and actual outcomes (e.g., Agnew 1992). The resulting negative emotions can, for instance, be anger or anxiety (e.g., Carmichael and Piquero 2004). Researchers assume that, when faced with strains, stressors, or stressful situations, perceptions of frustration and unfairness arise, which in turn are crucial mechanisms for the link between strain and dishonest behaviour (e.g., Agnew 1992; Agnew and White 1992; Freiburger et al. 2017).

Instead of being contradictory theories, researchers today often propose that both theories together may explain dishonesty, cheating, and deviant behaviour. Negative emotions and strain can influence how rational choices are interpreted, thus influencing individuals' cost–benefit calculations: For instance, negative emotions can reduce individuals' concerns of getting caught, thereby reducing the costs of potential dishonesty; negative emotions can also increase individuals' justifications and rationalizations for their dishonest behaviour (e.g., Carmichael and Piquero 2004; Fida et al. 2015, 2018). In line with this, negative emotions induced by stressors can also increase individuals' perceptions of the importance of potential benefits or the importance of rewards gained by their deceptive behaviour (e.g., Carmichael and Piquero 2004; Fida et al. 2015, 2018).

1.2 Direct and indirect effects of tests as difficult learning tasks on academic cheating

Notably, the above described negative consequences of desirable difficulties in the learning context (e.g., stress, anxiety, perceptions of unfairness) fit the just presented theories explaining dishonesty and academic cheating. Thus, a direct relation between tests as an incantation of desirable difficulties and academic cheating, as well as an indirect relation between tests, thereby inflicted negative consequences (e.g., stress, anxiety, perceptions of unfairness), and academic cheating can be assumed.

For instance, worries about doing well in school, getting good grades, teachers' evaluations, and about the own performance compared to the performance of peers were positively correlated to cheating (e.g., Anderman et al. 1998). Thus, students often cheat to increase their performance and to make a good impression on others (e.g., Franklyn-Stokes and Newstead 1995; Newstead et al. 1996; Wowra 2007). Fear of not being able to succeed, an inability of keeping up with the assignments, lower self-efficiency, and fear of failure were also linked to more academic cheating and often reported as reasons for (past) cheating (e.g., Finn and Frone 2004; McCabe 1992; Schab 1991; Whitley 1998). Notably, as described before, we suppose that tests as difficult learning tasks increase such perceptions of performing poorly and fear of failure because they are difficult, hard to solve, and because they reduce learners' illusion of competence and reduce their overconfidence.

Test anxiety, social anxiety, and general anxiety were also positively correlated to (past) academic cheating (e.g., Rost and Wild 1994; Whitley 1998; Wowra 2007). Stress, parental pressure, pressure for good grades, and pressure in general were also often found to be linked to cheating or were reported as reasons and incentives for such dishonest behaviour (e.g., Brimble and Stevenson-Clarke 2005; Davis et al. 1992; Schab 1991; Whitley 1998). A study by Steininger et al. (1964) further showed that the more negative a (test) situation was perceived, the more anxiety-provoking it was; or, the more a test was perceived as difficult, the more cheating was considered as justified and the more participants reported that they would cheat. Negative emotions due to stressors—and we suppose learning tests to be acute stressors—were further correlated to more moral disengagement in the work context, and in turn to more justifications for deceptive or counterproductive work behaviour (e.g., Fida et al. 2015). Such moral disengagement and justifications thus increased deceptive or counterproductive work behaviour (e.g., Fida et al. 2015), which could also apply to deception in the academic context. Notably, as described before, tests and difficult learning tasks were shown to increase such negative emotions and perceptions of stress, anxiety, and pressure (e.g., Hinze and Rapp 2014; O'Neil et al. 1969; Study 2, Wenzel and Reinhard 2019b).

Furthermore, students' perceptions of the course or assessments as (too) difficult increased academic misconduct (e.g., Brimble and Stevenson-Clarke 2005; Freiburger et al. 2017), and the difficulty of the course was sometimes described as one reason to justify, rationalize, or neutralize cheating behaviour (e.g., Haines et al. 1986). In line with this, higher workload was also linked to more cheating (e.g., McCabe 1992; Whitley 1998). Similarly, participants who thought they had

indulged in more effort in a task felt more entitled and felt that they had earned good outcomes, like higher grades, which in turn led to more moral justifications (e.g., Hoffman and Spitzer 1985). Notably, tests as incantations of desirable difficulties are, even by definition, difficult and they logically increase learners' effort and workload. Thus, we suppose that these findings should also apply to tests as difficult learning tasks.

Another study showed that the more a test situation was perceived as pressuring and as uncomfortable, that is, the more it was perceived as a high-pressure situation, the more unfair the testing tool was perceived (Leiner et al. 2018). Because most learners often believe easier and fluent learning to be extremely effective (e.g., Karpicke et al. 2009; Kornell et al. 2011), we assume that they should perceive difficult tasks like tests and the increased effort they require as unfair and generally as negative, especially when they are forced to use tests in their university courses. In turn, students often reported that they would cheat more and that cheating was more justified when they perceived their teachers, the teaching practices, or the assessments as unfair and their schools as extremely competitive (e.g., Brimble and Stevenson-Clarke 2005; Calabrese and Cochran 1990; Finn and Frone 2004; LaBeff et al. 1990; McCabe 1992; Olafson et al. 2013; Whitley 1998). Fittingly, people who generally thought they were being treated unfairly were more inclined toward dishonesty (e.g., Houser et al. 2012) and perceived inequity was linked to more deceptive behaviour (e.g., Greenberg 1990).

1.3 The present research

In summary, the just described theoretical assumptions and the fitting empirical findings indicate that the application of tests as difficult learning tasks can directly or indirectly (via increasing negative consequences like perceptions of stress, anxiety, or feelings of unfairness) lead to more academic cheating. In more detail, difficult learning tests were argued to result in negative consequences like more stress, more negative perceptions and emotions, or more feelings of unfairness. These negative consequences were in turn often found to be linked to more cheating, more intentions to cheat, and to more justifications for cheating. Hence, the present work was conducted to test these theoretically derived direct and indirect effects of tests as difficult learning tasks on students' academic cheating.

Notably, there are to our knowledge neither studies exploring academic cheating as a result of tests as difficult learning tasks nor studies exploring academic cheating as a result of negative consequences like stress perceptions and negative situation evaluations caused by tests. Most of the existing studies regarding desirable difficulties focused on individual abilities or external factors serving as moderators or requirements for the described beneficial effects (see e.g., Adesope et al. 2017; Dobson and Linderholm 2015; Rowland 2014). However, the main focus was seldom on further negative consequences beyond reduced or restricted learning success and seldom on further triggered behaviour like cheating. We nonetheless argue that it is important to focus on these (new) assumptions because academic cheating can be seen as a widespread and problematic behaviour, even though students

themselves normally perceive cheating during an exam as having rather light consequences (because it is perceived as not directly harming others; e.g., Brimble and Stevenson-Clarke 2005; Marksteiner et al. 2013). For instance, due to cheating on a test, teachers cannot accurately grade students and can therefore not appropriately support their learning processes or help them to increase their skills (e.g., Reinhard et al. 2011). Students who enhance their performance through cheating can also gain an unfair and undeserved advantage compared to others, distort the performance succession in a class, increase competition, trigger peer cheating, and even normalize dishonest behaviour (e.g., Carrell et al. 2008; Fida et al. 2018; Gino et al. 2009; McCabe et al. 2001; Paternoster et al. 2013). Dishonesty in an academic setting is also often linked to further dishonesty in later workplaces (e.g., Nonis and Swift 2001). Due to these negative impacts of academic cheating and due to the lack of previous work, we think that it is relevant to investigate if the application of tests as difficult learning tasks directly or indirectly increases the probability of cheating before advising the usage of such learning tasks in universities. Hence, the present study uniquely contributes to the literature on desirable difficulties and to the literature on cheating behaviour.

To measure dishonest behaviour, researchers often use scenarios because these are assumed to accurately mirror emotions, intentions, and behaviours of individuals in different situations (e.g., Agnew 1992; Carmichael and Piquero 2004; Shu et al. 2011). Thus, we conducted an online study with the learning scenario condition (divided in one reading control scenario condition and two test scenario conditions) as the between-subjects variable. We further assessed individuals' negative evaluations of the learning scenarios as well as their stress perceptions in such imagined situations as two potential mediators. Self-reported likelihoods of hypothetical cheating and justifications for cheating served as our dependent variables.

1.4 Hypotheses

Due to the argumentations presented above, we assume the following hypotheses (see Fig. 1 for a conceptual diagram of the assumed relations): We suppose that both learning scenario conditions with tests lead to more negative evaluations of the learning situations (*Hypothesis 1*) and to higher stress perceptions (*Hypothesis 2*) than the reading control learning scenario condition. Both learning scenario conditions with tests are further assumed to directly lead to higher likelihoods of hypothetical cheating than the reading control learning scenario condition (*Hypothesis 3*). The negative evaluations of the learning situations are also hypothesized to be positively correlated to likelihoods of hypothetical cheating (*Hypothesis 4*). In line with this, stress perceptions are further assumed to be positively correlated to likelihoods of hypothetical cheating (*Hypothesis 5*). Moreover, we assume that both learning scenario conditions with tests directly lead to higher justifications for hypothetical cheating than the reading control learning scenario condition (*Hypothesis 6*). The negative evaluations of the learning situations are also hypothesized to be positively correlated to justifications for cheating (*Hypothesis 7*). In line with this, stress perceptions are assumed to be positively correlated to justifications for

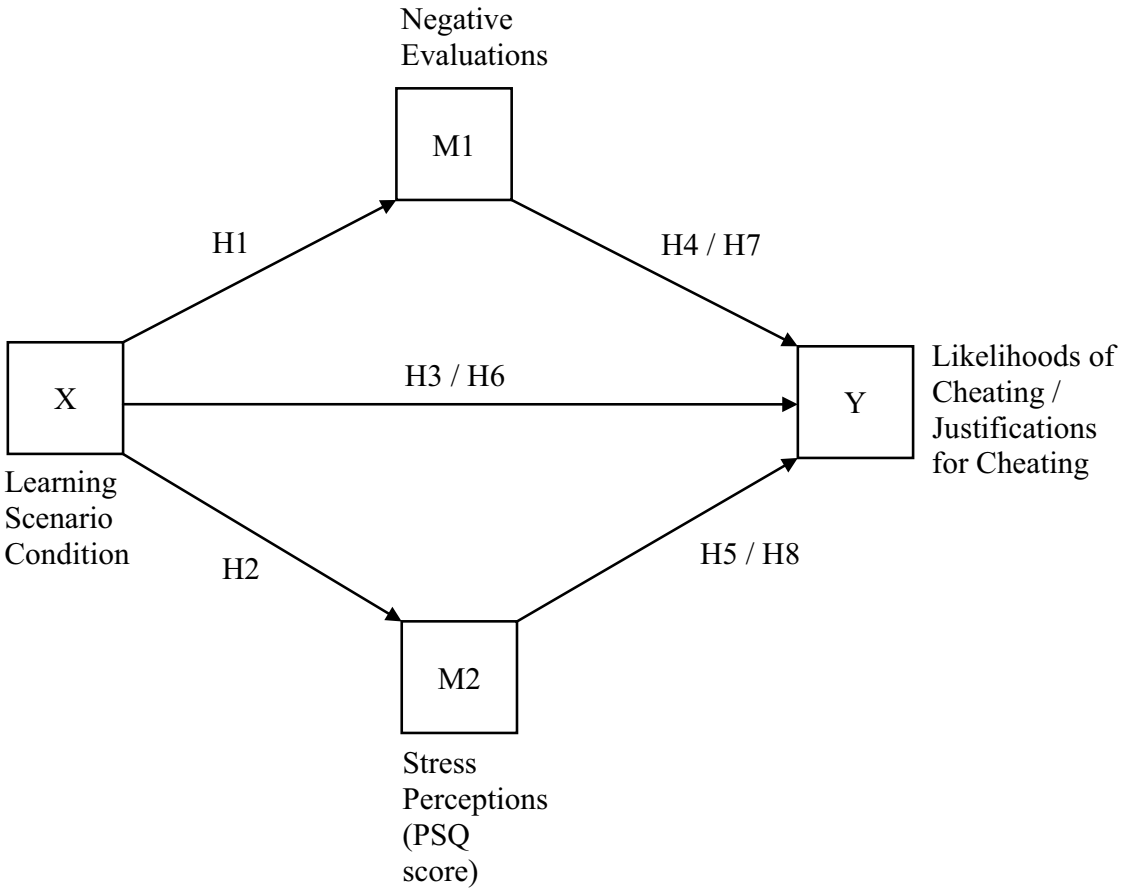


Fig. 1 Conceptual diagram of the assumed hypotheses. *Notes*. The learning scenario condition (X) includes a reading control scenario, a test with private results learning scenario, and a test with public results learning scenario

cheating (*Hypothesis 8*). Thus, apart from direct effects of the learning scenario condition on likelihoods of hypothetical cheating and on justifications for cheating, indirect effects via increases of the negative evaluations of the learning situations and via increases of stress perceptions are also assumed.

2 Methods

2.1 Participants

Power was set to .95 and sample size was calculated to detect a small to medium effect ($f = .20$). Using G*Power (Faul et al. 2009), a power analysis revealed a required sample size of $N = 390$ to detect a significant effect (alpha level of .05), given there is a true effect. To test our hypotheses, we recruited an American online sample consisting of 458 participants, 53 of whom were excluded because they answered at least one of three attention-check questions incorrectly. Thus, our final sample consisted of $N = 405$ participants from MTurk ($M_{age} = 25.72$, $SD_{age} = 6.65$, range = 18–62, 48.4% female, 97.3% English native speakers,

all college or university students). Each participant was randomly assigned to one of the three learning scenario conditions: either the test with public results learning scenario condition ($n = 129$), the test with private results learning scenario condition ($n = 136$), or the reading control learning scenario condition ($n = 140$). Before starting the experiment, all participants had to provide their approval through reading and then agreeing to an informed consent (stating that they knew that their participation was completely voluntary and that they could withdraw at any time without explanation); participants also confirmed that they were at least 18 years old. The study was conducted in accordance with the Ethical Guidelines of the DGPs as well as the APA, and the project was approved by the Ethics Committee affiliated with the funding source. Participants received .60$ for their participation.

2.2 Procedure and measures

The present work was conducted together with another study (concerning desirable difficulties, trait variables potentially linked to perceptions of such difficult learning situations, and by desirable difficulties caused stress experiences; Study 1, Wenzel and Reinhard 2019b). Our dependent variables assessing likelihoods of hypothetical cheating and justifications for cheating were assessed at the end of this other study.

At the beginning, participants read brief details about the study and then answered some questions regarding demographics, e.g., age, gender, and native language. Thereafter, different trait variables (e.g., trait test anxiety and trait stress) were assessed solely for the other study (Study 1, Wenzel and Reinhard 2019b; academic self-concept: Dickhäuser et al. 2002; PAF-E: Hoferichter et al. 2016; PSS: Cohen et al. 1983; SSS: Reeder et al. 1973). Although these dispositional variables may be related to the dependent variables of the present work, they will not be included in the analyses because dispositional variables were—unlike the direct and indirect effects of the learning situations—not the focus of the present study.

Participants were then randomly assigned to one of the three learning scenario conditions. As an example, the test with public results learning scenario condition, including the instructions, reads as follows:

> This is a potential scenario that could happen in your daily life as a student. We would like to ask you to transport yourself in the situation, and to imagine it as strongly as you can. Imagine that you are a student in college and have lots of exams to write. During one of your majors your professor tries to increase your and your fellow students learning success, and enhance your chance to pass the exam. Therefore, half an hour before the end of every session you write an ungraded test, and answer multiple questions concerning the content of that session. Once the half an hour is up you can go home. Shortly following every session all students receive an e-mail with the matriculation numbers of everyone, and their test results, ranking from best to worst.

In the test with private results learning scenario condition, in which stakes should be perceived as even lower, participants read a slightly different scenario and were instructed to imagine that each student received the test results individually via

e-mail. In contrast, in the reading control learning scenario condition, the imagined process was that the professor would hand the students a summary of all relevant materials to read. See "Appendix A" for all three learning scenarios.

To follow, participants answered questions concerning their perceptions and evaluations of the imagined learning scenario, e.g., regarding difficulty, unfairness, inappropriateness, anger, or injustice. This concluded in an overall *negative evaluations of the learning situations* score using ten items ($\alpha = .89$; e.g., *How (un) just did you find the described and imagined way of learning in the situation?,* one (*extremely unjust*)—seven (*extremely just*)) on a seven-point Likert-like scale from one (lower scores) to seven (higher scores). Some of the items were reverse coded (e.g., participants were asked how fair they thought the learning in the scenario was). See "Appendix A" for a full list of all items, information about which items were recoded, and the complete scale labelling. We also added three—later not analysed—positive control items (e.g., asking for the perceived helpfulness or successfulness of such learning tasks) so that it was not completely clear that we wanted to assess an overall negative evaluations score. We added these positive control items because we wanted to avoid being too obvious, being potentially suggestive, or to unintentionally influence participants' later responses. Participants were also asked about their situational *stress perceptions* in such an imagined learning scenario using the Perceived Stress Questionnaire (PSQ; Levenstein et al. 1993) that consists of 30 items ($\alpha = .95$; e.g., *You feel tense*) on a four-point Likert-like scale from one (*almost never*) to four (*usually*).

Subsequently, participants were told to again put themselves in the aforementioned scenario and to read the following statement regarding a hypothetical examination:

> While preparing for the exam you took little notes and prepared a crib sheet you only wanted to use for your learning. Now imagine that you are in class with your fellow students writing the exam. Thinking about the answer to question number one you suddenly realize that the crib sheet you used to practice is still in your pocket.

Participants were then asked how likely it was for them to use the crib sheet to cheat on the exam (cheating item 1: *likelihoods own spontaneous cheating*) and how justifiable that was (cheating item 2: *justifications own spontaneous cheating*). Then, participants had to rate how likely it was for someone else to use the crib sheet to cheat on the exam (cheating item 3: *likelihoods others' spontaneous cheating*) and how justifiable that was (cheating item 4: *justifications others' spontaneous cheating*). Participants were then asked how likely it was for them to intentionally prepare a cheat sheet with the aim to use it during the exam (cheating item 5: *likelihoods own prepared cheating*) and how justifiable that was (cheating item 6: *justifications own prepared cheating*). They also reported how likely it was for someone else to intentionally prepare a cheat sheet with the aim to cheat during the exam (cheating item 7: *likelihoods others' prepared cheating*) and how justifiable that was (cheating item 8: *justifications others' prepared cheating*). These eight *cheating items*—four likelihoods items and four justifications items—were answered on a seven-point Likert-like scale from one (*not likely at all/not justifiable at all*) to seven (*extremely*

likely/extremely justifiable). See "Appendix A" for a full list of the items. In line with previous research (see e.g., Greene and Saxe 1992; Messick et al. 1985; Shu et al. 2011), we added items distinguishing between likelihoods and justifications for own hypothetical cheating behaviour and likelihoods and justifications for hypothetical cheating behaviour of other people. Because these cheating items were newly created for our study, we ran factorial analyses to test the underlying number of factors before testing our hypotheses. Regarding the four likelihoods of cheating items, the factor analysis yielded two factors: Factor 1 consisted of the two items regarding the likelihoods of own cheating (average score of the two items: *likelihoods own cheating,* $\alpha = .86$) and factor 2 consisted of the two items regarding the likelihoods of others' cheating (average score of the two items: *likelihoods others' cheating,* $\alpha = .84$). The second factor analysis was conducted with the four justification for cheating items and resulted in one factor (average score across the four items: *justifications for cheating,* $\alpha = .95$). A detailed description of the two factor analyses and the respective tables depicting the loadings of the factor analyses are available in "Appendix B".

In the end, we measured general control variables (e.g., if participants had really imagined the read scenarios, if they understood the text, or how strongly they were able to put themselves in the learning scenarios). For instance, one item reads, *Did you understand the described scenario?*, and it was rated from one (*No, not at all*) to seven (*Yes, completely*). See "Appendix A" for a list of these items. We also included manipulation check questions regarding cheating (e.g., how important grades are for the participants, if they think they can improve their results through cheating, how likely it was to get caught in the imagined scenario, how likeable they rated the imagined lecturer in the scenario, and if they held negative or positive attitudes towards cheating). For instance, one item reads, *How likeable would you rate your professor?*, rated from one (*absolutely unlikeable*) to seven (*extremely likeable*). See "Appendix A" for a list of these items. These manipulation check questions were included to test for differences among participants in the three learning scenario conditions.

2.3 Statistical analyses

To test our hypotheses, we conducted three mediation analyses using PROCESS (Hayes 2018; model 4). Due to the factor analysis that yielded two factors for likelihoods of hypothetical cheating—one for own cheating behaviour and one for others' cheating behaviour—we conducted two analyses to test the hypotheses that concern likelihoods of hypothetical cheating (e.g., predicting the influence of the learning scenario condition on likelihoods of hypothetical cheating as well as linkages between negative evaluations of learning situations and stress perceptions with likelihoods of cheating; Hypotheses 3, 4, and 5).

The first mediation analysis used likelihoods of own cheating (testing Hypotheses 1, 2, 3a, 4a, and 5a), the second mediation analysis used likelihoods of others' cheating (testing Hypotheses 3b, 4b, and 5b), and the third mediation analysis used justifications for cheating (testing Hypotheses 6, 7, and 8) as the respective dependent

variable. All three mediation analyses used the learning scenario condition as the independent variable and participants' negative evaluations of the learning situations as well as participants' stress perceptions as two potential mediators. The learning scenario condition was dummy coded (X1: 1 = tests with private results learning scenario condition; X2: 1 = tests with public results learning scenario condition; reference category: reading control scenario condition). The mediator variables were z-standardized. To avoid unnecessary repetitions, only the description of the findings of the first mediation analysis will include the influence of the learning scenario condition on the two mediators, thus, on the negative evaluations of the learning situations and on participants' stress perceptions (testing Hypotheses 1 and 2).

3 Results

Neither participants' gender distribution nor their age differed among the three learning scenario conditions (both $ps \geq .230$). The general control variables and the manipulation check questions regarding cheating also did not differ among the three learning scenario conditions (all $ps \geq .091$). Only participants in the test with public results learning scenario condition rated the lecturer as more dislikeable than participants in the other two learning scenario conditions (both $ps \leq .001$).

The descriptive statistics of the negative evaluations of the learning situations, stress perceptions indicated by PSQ scores, likelihoods of own cheating, likelihoods of others' cheating, and justifications for cheating are presented in Table 1. Notably, likelihoods of others' hypothetical cheating were rated as significantly higher than likelihoods of own hypothetical cheating ($p < .001$).

The correlations among participants' negative evaluations of the learning situations, stress perceptions (PSQ), likelihoods of own hypothetical cheating, likelihoods of others' hypothetical cheating, and justifications for cheating are depicted in Table 2. Notably, the negative evaluations of the learning situations were significantly correlated to participants' likelihoods of own cheating ($r = .19$, $p < .001$) and to participants' justifications for cheating ($r = .16$, $p < .001$). The PSQ scores indicating stress perceptions were significantly correlated to likelihoods of others' cheating ($r = .14$, $p = .006$).

Table 1 Descriptive statistics of the negative evaluations of the learning situations, stress perceptions (PSQ), likelihoods of own cheating, likelihoods of others' cheating, and justifications for cheating

Variables	M	SD	Range
Negative evaluations of the learning situations	3.58	1.15	1.00–7.00
Stress perceptions (PSQ)	2.21	.63	1.00–4.00
Likelihoods own cheating	2.29	1.55	1.00–7.00
Likelihoods others' cheating	4.21	1.46	1.00–7.00
Justifications for cheating	2.32	1.55	1.00–7.00

$N = 405$

Table 2 Correlations among the negative evaluations of the learning situations, stress perceptions (PSQ), likelihoods own cheating, likelihoods others' cheating, and justifications for cheating

	1.	2.	3.	4.	5.
1. Negative evaluations of the learning situations	1				
2. Stress perceptions (PSQ)	.50**	1			
3. Likelihoods own cheating	.19**	.08^{+}	1		
4. Likelihoods others' cheating	.04	.14**	.36**	1	
5. Justifications for cheating	.16**	.08	.86**	.33**	1

$^{+}p < .10$, $^{*}p < .05$, $^{**}p < .01$, Two-tailed. $N = 405$

3.1 Likelihoods own hypothetical cheating (Hypotheses 1, 2, 3a, 4a, and 5a)

Results of the first mediation analysis (see Fig. 2) showed that the learning scenario condition significantly predicted participants' negative evaluations of the learning situations (path a), X1: $B = .38$, $SE = .12$, $t(402) = 3.27$, $p = .001$; X2: $B = .79$, $SE = .12$, $t(402) = 6.66$, $p < .001$. In turn, the negative evaluations of the learning

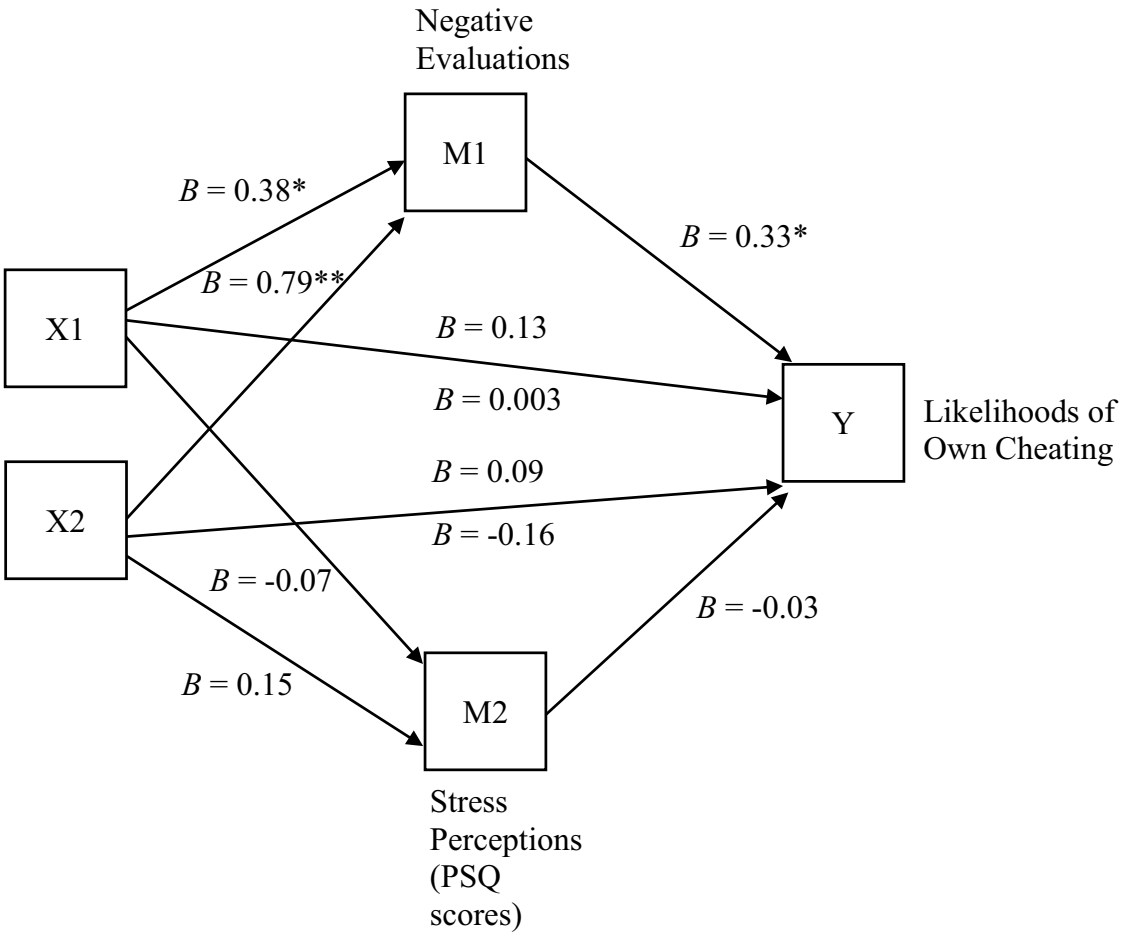


Fig. 2 First mediation analysis predicting likelihoods of own cheating. *Notes*. $^{*}p < .05$, $^{**}p < .01$. The learning scenario condition was dummy coded (X1: 1 = tests with private results learning scenario condition; X2: 1 = tests with public results learning scenario condition; reference category: reading control scenario condition)

situations predicted participants' likelihoods of own hypothetical cheating (path b), $B = .33$, $SE = .09$, $t(400) = 3.61$, $p = .003$. The learning scenario condition did not significantly predict participants' stress perceptions indicated by their PSQ scores (path a), X1: $B = -.07$, $SE = .12$, $t(402) = -.61$, $p = .545$; X2: $B = .15$, $SE = .12$, $t(402) = 1.24$, $p = .216$. PSQ scores were also not linked to participants' likelihoods of own cheating (path b), $B = -.03$, $SE = .09$, $t(400) = .02$, $p = .706$. There was no significant direct effect (path c') of the learning scenario condition on likelihoods of own cheating, X1: $B = .003$, $SE = .19$, $t(400) = .02$, $p = .986$; X2: $B = -.16$, $SE = .20$, $t(400) = -.82$, $p = .411$. There was also no significant total effect (path c) of the learning scenario condition on likelihoods of own cheating, X1: $B = .13$, $SE = .19$, $t(402) = .71$, $p = .481$; X2: $B = .09$, $SE = .19$, $t(402) = .49$, $p = .626$. However, the results yielded significant indirect effects of the learning scenario condition via the negative evaluations of the learning situations on likelihoods of own hypothetical cheating (path a x path b), X1: $B = .13$, 95% CI [.037, 242]; X2: $B = .26$, 95% CI [.113, .423]. There were no indirect effects of the learning scenario condition via the PSQ scores, X1: $B = .002$, 95% CI [−.023, .030]; X2: $B = -.01$, 95% CI [−.050, .025].

These findings supported Hypothesis 1: Both learning scenarios including tests were, as assumed, evaluated more negatively than the reading control learning scenario. These negative evaluations included, for instance, higher perceptions of unfairness, strenuousness, and injustice, as well as higher feelings of anger. Unexpectedly, the learning scenario condition neither influenced participants' stress perceptions nor likelihoods of participants' own hypothetical cheating. Thus, Hypotheses 2 and 3a were not supported. In line with our assumptions—supporting Hypothesis 4a—negative evaluations of the learning situations were significantly and positively correlated to participants' own hypothetical cheating, indicating that higher negative evaluations were linked to higher likelihoods of own cheating. This indirect effect of the learning scenario condition on likelihoods of own hypothetical cheating (via increased negative evaluations of the learning situations) was significant. Hence, negative evaluations of the learning situations had a mediating effect. Contrary to Hypothesis 5a, stress perceptions were not significantly correlated to participants' likelihoods of own hypothetical cheating.

3.2 Likelihoods of others' hypothetical cheating (Hypotheses 3b, 4b, and 5b)

Results of the second mediation analysis (see Fig. 3) showed that the negative evaluations of the learning situations did not predict likelihoods of others' hypothetical cheating (path b), $B = -.07$, $SE = .09$, $t(400) = -.80$, $p = 425$. The PSQ score was, however, linked to participants' likelihoods of others' cheating (path b), $B = .23$, $SE = .08$, $t\ (400) = 2.74$, $p = .007$. There was again no significant direct effect (path c') of the learning scenario condition on likelihoods of others' cheating, X1: $B = .01$, $SE = .18$, $t(400) = .07$, $p = .946$; X2: $B = .11$, $SE = .19$, $t(400) = .58$, $p = .560$. There was also no significant total effect (path c) of the learning scenario condition on likelihoods of others' cheating, X1: $B = -.03$, $SE = .18$, $t(402) = -.18$, $p = .859$; X2: $B = .09$, $SE = .18$, $t(402) = .50$, $p = .615$.

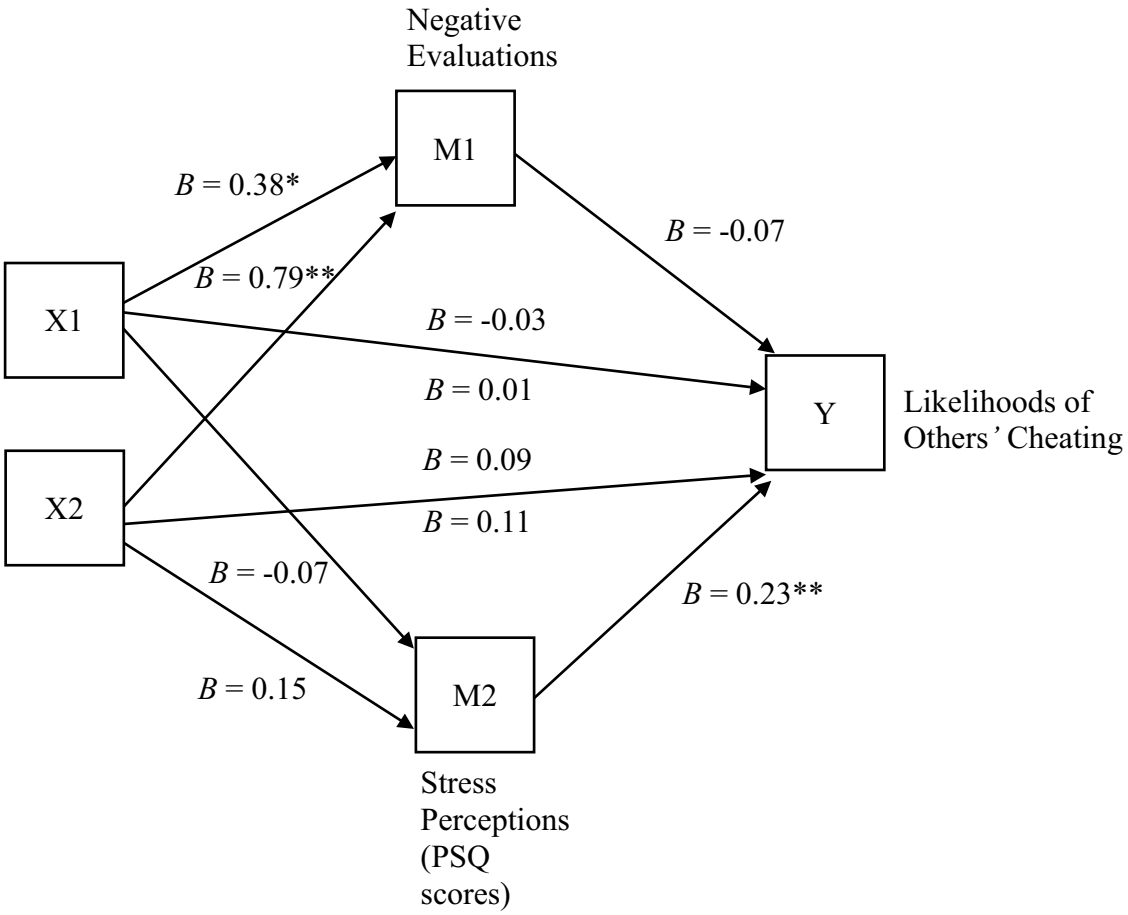


Fig. 3 Second mediation analysis predicting likelihoods of others' cheating. *Notes*. *$p < .05$, **$p < .01$. The learning scenario condition was dummy coded (X1: 1 = tests with private results learning scenario condition; X2: 1 = tests with public results learning scenario condition; reference category: reading control scenario condition)

Additionally, the findings yielded no significant indirect effects of the learning scenario condition via the negative evaluations of the learning situations on likelihoods of others' cheating (path a x path b), X1: $B = -.03$, 95% CI [−.105, .044]; X2: $B = -.06$, 95% CI [−.197, .085]. There were also no indirect effects of the learning scenario via the PSQ scores, X1: $B = -.02$, 95% CI [−.086, .039]; X2: $B = .04$, 95% CI [−.021, .105].

Unexpectedly, the learning scenario condition did not influence likelihoods of others' hypothetical cheating. Thus, Hypothesis 3b could not be supported, indicating that the learning scenario had no effect on individuals' ratings of the probability of others' cheating in a hypothetical examination. Contrary to Hypothesis 4b, negative evaluations of the learning situations were not significantly linked to others' hypothetical cheating. Participants' stress perceptions were, however, significantly and positively correlated to likelihoods of others' hypothetical cheating, thus, supporting Hypothesis 5b. This indicated that higher stress perceptions were linked to higher ratings regarding likelihoods of others' hypothetical cheating behaviour. There were no indirect effects.

3.3 Justifications for hypothetical cheating (Hypotheses 6, 7, and 8)

Results of the third mediation analysis (see Fig. 4) showed that the negative evaluations of the learning situations significantly predicted justifications for cheating (path b), $B = .24$, $SE = .09$, $t\ (400) = 2.56$, $p = .011$. The PSQ scores indicating stress perceptions were not linked to justifications for cheating (path b), $B = -.003$, $SE = .09$, $t\ (400) = -.38$, $p = .970$. There was again no significant direct effect (path c') of the learning scenario condition on justifications for cheating, X1: $B = .02$, $SE = .19$, $t(400) = .13$, $p = .901$; X2: $B = .07$, $SE = .20$, $t(400) = .33$, $p = .743$. There was also no significant total effect (path c) of the learning scenario condition on justifications for cheating, X1: $B = .11$, $SE = .19$, $t(402) = .61$, $p = .542$; X2: $B = .25$, $SE = .19$, $t(402) = 1.33$, $p = .186$. However, the findings yielded significant indirect effects of the learning scenario condition via negative evaluations of the learning situations on justifications for cheating (path a x path b), X1: $B = .09$, 95% CI [.016, .188]; X2: $B = .19$, 95% CI [.044, .341]. There were no indirect effects of the learning scenario condition via the

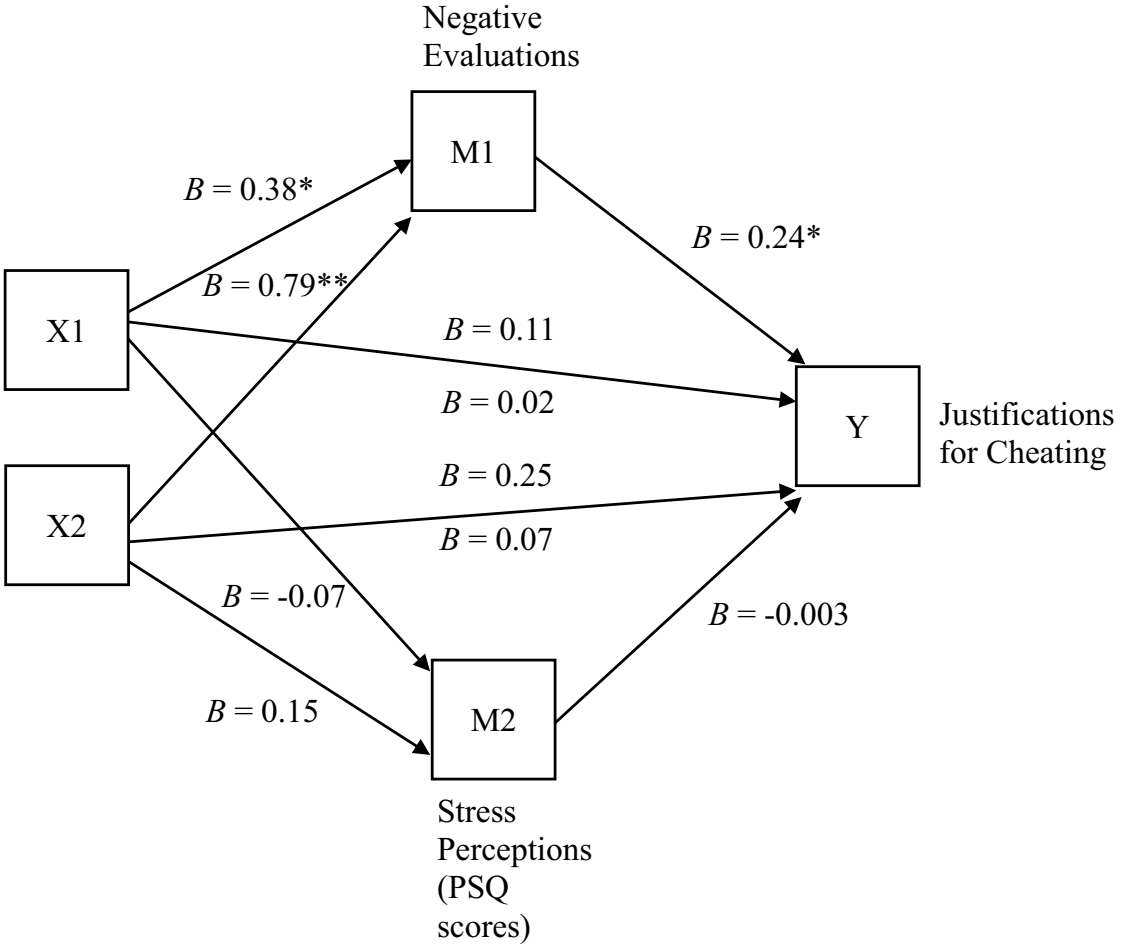


Fig. 4 Third mediation analysis predicting justifications for cheating. *Notes*. *$p < .05$, **$p < .01$. The learning scenario condition was dummy coded (X1: 1 = tests with private results learning scenario condition; X2: 1 = tests with public results learning scenario condition; reference category: reading control scenario condition)

PSQ scores, X1: $B < .001$, 95% CI [− .029, .026]; X2: $B = -.001$, 95% CI [− .042, .031].

Contrary to Hypothesis 6, the learning scenario condition did not influence participants' justifications for hypothetical cheating. Thus, participants' ratings of justifications for hypothetical cheating were not dependent on whether participants had read scenarios including tests or including reading tasks. Negative evaluations of the learning situations were significantly and positively correlated to justifications for hypothetical cheating. This supported Hypothesis 7 and indicated that higher negative evaluations of the learning situations were linked to later higher justifications for cheating in the university context. The indirect effect of the learning scenario condition on justifications for hypothetical cheating (via increased negative evaluations of the learning situations) was also significant. Hence, negative evaluations of the learning situations had a mediating effect. Participants' stress perceptions were not correlated to justifications for hypothetical cheating. Thus, Hypothesis 8 was not supported.

4 Discussion

The aim of the present work was to test linkages among tests as difficult learning tasks, possible negative consequences of such difficult learning tasks like negative evaluations or stress perceptions, and hypothetical academic cheating. We assumed that learning scenarios including tests, as opposed to a control learning scenario including reading, directly and indirectly, lead to higher likelihoods of own and others' hypothetical cheating, as well as to higher justifications for such hypothetical cheating. The indirect effects should arise via increased negative evaluations of the learning situations and via increased stress perceptions due to the difficult test scenarios. Although ample research has focused on the application and effectiveness of tests as incantations of desirable difficulties (e.g., regarding potential moderators or boundary conditions; e.g., Adesope et al. 2017; Rowland 2014) and although previous studies showed that academic cheating has an abundance of negative impacts (e.g., regarding contagion effects of dishonesty through peers, relations between academic and workplace dishonesty, or validity of assessments and grading; e.g., Carrell et al. 2008; Gino et al. 2009; Nonis and Swift 2001; Reinhard et al. 2011), no research has—to our knowledge—previously tested our assumptions and hypotheses. Thus, our work using hypothetical scenarios uniquely contributes to the existing literature regarding cheating in the academic context and to the existing literature regarding tests as difficult learning tasks.

Our findings showed that although the learning scenario condition had neither direct effects on likelihoods of own and others' hypothetical cheating nor on justifications for cheating, it nonetheless indirectly affected likelihoods of own cheating and justifications for cheating through increasing participants' negative evaluations of the learning situations. Both imagined learning scenarios including tests were evaluated as significantly more negative than the learning control scenario including reading. These negative evaluations of the learning situations were in turn positively correlated with likelihoods of own hypothetical cheating and with justifications for

cheating, whereas participants' self-reported stress perceptions were only positively correlated to likelihoods of others' hypothetical cheating. In general, the cheating items had rather low mean scores, whereas likelihoods of others' hypothetical cheating were rated as significantly higher than likelihoods of own hypothetical cheating. This finding is in line with previous work showing that students often report lower frequencies of academic cheating compared to the amount that lecturers observed, and that students also report lower frequencies of their own dishonest behaviour compared to dishonest behaviour of their peers (e.g., Greene and Saxe 1992; Naghdipour and Emeagwali 2013). Additionally, students often perceive their own dishonest behaviour as less condemnable and less serious than the dishonest behaviour of their peers and generally believe that they are fairer than others (e.g., Greene and Saxe 1992; Messick et al. 1985). Thus, it could be that individuals underreport their own cheating behaviour (even in anonymous settings), likely because of the importance and value of norms like honesty and trustworthiness, the urge to maintain a positive self-concept, and the underlying social undesirability of dishonesty (e.g., Geißler et al. 2013; Mazar et al. 2008). Our factor analysis that yielded one factor regarding likelihoods of own hypothetical cheating and a second factor regarding likelihoods of others' hypothetical cheating further supported these findings. Interestingly, our factor analysis revealed only one factor underlying the four justifications for cheating items. Thus, our participants did not distinguish between justifications for own hypothetical cheating behaviour and justifications for others' hypothetical cheating behaviour (contrary to previously found differences between justifications for own and others' dishonesty, see e.g., Shu et al. 2011). In general, the rather low mean score of the justifications for cheating variable further indicates that participants rated hypothetical cheating in the presented scenarios as not justifiable, thus deeming academic cheating as ethically wrong. An explanation for the observed single factor could be that individuals normally try to maintain a positive self-concept and try to feel good or moral even when they cheat (e.g., Mazar et al. 2008): Therefore, they often compare their own behaviour with others' behaviour and, for instance, often believe that others cheated even more—and more severely—than they did (see e.g., Greene and Saxe 1992). This social comparison should, however, only increase individuals' perceptions of themselves as a better or more moral individual compared to others, if the justifications for their own and for others' behaviours are identically low-, because only then should the higher frequencies of others' dishonest behaviour compared to individuals' own less frequent dishonesty increase individuals' self-esteem and their moral self-concept. Moreover, it could also be possible that justifications for own cheating behaviour and justifications for others' cheating behaviour only significantly differ if the justifications ratings were rendered after individuals indulged in actual dishonest behaviour and not just in response to imagined hypothetical cheating.

Notably, our results were obtained even though participants did not really engage in an actual learning activity; they did not really take an exam with actual consequences for their everyday courses, but simply read and imagined scenarios and only self-reported hypothetical behaviour. Nonetheless, even such minimalistic operationalizations yielded significant effects. Thus, this indicates that actual learning in university settings, with real incentives to do well and with actual examinations

including opportunities of actual cheating behaviour, should lead to even stronger effects.

Our results partly fit the in the beginning described theoretical and empirical argumentations regarding negative consequences of desirable difficulties because the scenarios including tests were actually evaluated more negatively than the reading control scenario (e.g., Hinze and Rapp 2014; O'Neil et al. 1969). The observed indirect effects of learning scenarios with tests on own hypothetical cheating behaviour and justifications—via increased negative evaluations of the situations—were also in line with the in the Introduction presented theoretical and empirical argumentations regarding the emergence of cheating and dishonesty in academic contexts (e.g., Brimble and Stevenson-Clarke 2005; Steininger et al. 1964; Whitley 1998; Wowra 2007). Contrary to our assumptions and to literature described in the Introduction, there were neither effects of the learning scenario condition on participants' stress perceptions nor direct effects of the learning scenario condition on the cheating variables. This could be due to our operationalizations and the application of hypothetical scenarios: It is possible that our scenarios were not strong enough to elicit actual affective responses as well as hypothetical cheating behaviour in only imagined situations (see also our discussion of limitations below).

Although not all our hypotheses were supported, it is still important to highlight that tests as difficult learning tasks can, at least indirectly and in scenarios, influence hypothetical cheating behaviour. Hence, lecturers thinking about applying tests as difficult learning tasks in their university courses should keep in mind that these can result in negative evaluations of the situations and can, indirectly, also result in increased likelihoods of cheating or justifications for cheating. Still, due to the explorative character of our work and because this is to our knowledge the first study testing possible effects of tests as difficult learning tasks on cheating, we suppose that it is too early for stating implications like advising against the usage of tests and desirable difficulties. Nonetheless, our work sheds light on this problematic issue, offering a valuable contributing to the literature regarding desirable difficulties as well as cheating.

4.1 Limitations and future research

There are also limitations of our study that we care to discuss. This includes, for instance, the applied learning scenarios: Although scenarios are often used in studies focusing on cheating behaviour (e.g., Agnew 1992; Carmichael and Piquero 2004; Shu et al. 2011), it is possible that the learning scenarios had no effects because they were too short, not detailed enough, framed as positive, or too low-stake. Still, we intentionally designed them to be preferably short, generalizable (e.g., regarding varying study paths or courses), and minimalistic (e.g., so as not to be suggestive or influencing). We additionally wanted to inquire if effects would arise even when using such simple operationalizations. However, the scenarios may further have been unable to adequately describe and convey the increased effort, difficulty, and cognitive processing triggered by desirable difficulties. The same applies to the short description of the hypothetical exam at the end of the semester, which also

could have been too short, too undifferentiated, or not detailed enough, because the short scenario did not actually describe features of the examination situation (e.g., regarding the importance of the exam, the existence of peers, or the topic of the exam). This could have reduced the transportability and imaginability of the scenario. Although our intention was to not prime or suggest responses due to more detailed descriptions of the hypothetical examination (e.g., by describing opportunities to cheat or the difficulty of the exam), it is possible that more details concerning the examination situation would have made the scenario more comprehensible, more realistic, and more transferable to participants' actual experiences and everyday lives. We may have then been unable to control how participants actually imagined the examination situation, which might have resulted in confounding variables that, in turn, could have influenced participants' answers. Another limitation is that regarding the negative evaluations of the learning situations and the cheating variables, we only observed correlations. Future work should also test causal relations. To do this, future studies could, for instance, directly manipulate the evaluations of the learning situations described in the scenarios, so that the test scenarios as well as the reading control scenario are respectively described as positive, negative, and neutral. This would make it possible to explore whether all conditions including test scenarios lead to higher hypothetical cheating and justifications, or whether only those scenarios that were described as negative would increase hypothetical cheating and justifications for cheating.

In line with the novelty of our research questions and their unique contributions to the cheating and education literature, one of the best aspects of the present work is that it is surely stimulating for further research. For instance, future studies could try to optimize our operationalizations, thus solving the limitations mentioned above, and generally try to replicate our findings using different samples (e.g., students from different countries), different desirable difficulties (e.g., generation or disfluency), or different negative consequences (e.g., negative affect, fear of failure, or feelings of pressure). More explicitly, future studies could also be conducted in laboratory settings or in actual classrooms, applying a real learning phase including actually learned information, so that real—and not only hypothetical—cheating behaviour can be observed. Moreover, future online studies should test our assumptions using different and more detailed scenarios that more adequately describe the learning situation, the learning materials, and the difficulty of the learning tasks. The description of the examination should also be longer and more detailed, for example describing the procedure of the exam, the applied questions, the presence of peers or lecturers, and precautions against cheating more realistic. We also solely presented the usage of cheat sheets in examinations as the incantation of cheating behaviour; however, a far wider range of such behaviour does exist and should also be examined (e.g., inappropriate collaboration during exams or plagiarism). Additionally, until now, we focused completely on situational variables but not on individual variables, whereas previous studies showed that multiple trait variables, individual characteristics, and individual differences (e.g., cognitive abilities, conscientiousness, learning-goal orientations, self-control, or self-efficacy) are simultaneously influential for (difficult) learning (e.g., for perceptions or effectiveness's) and for cheating behaviour and dishonesty (e.g., Bertrams and Englert 2014; de Bruin and Rudnick 2007;

Doménech-Betoret et al. 2017; Finn and Frone 2004; Giluk and Postlethwaite 2015; Ikeda et al. 2015; Koul 2012; Marcela 2015; Paulhus and Dubois 2015; Schunk 1996; Wenzel and Reinhard 2019a; Yu et al. 2017; see also "Appendix B" regarding correlations among our dependent variables and the assessed but not analysed trait variables). Thus, we argue that it is beneficial for future work to include the assessment of individual differences. Lastly, future research should of course also focus on reducing such direct and indirect negative consequences of tests as difficult learning tasks. Lecturers could, for instance, thoroughly explain the benefits of difficult learning to their students, reward them for their efforts, frame the difficulties as even more positive and low-stake, and adapt the difficulty of the tasks so that they are difficult enough to elicit beneficial effects but are not too difficult or overwhelming.

4.2 Conclusion

Summarizing, the present work shows that the application of tests as an incantation of desirable difficulties in the university context—although normally beneficial for long-term learning—can result in negative side effects: Learning scenarios including tests, in contrast to a reading control scenario, indirectly increased likelihoods of own hypothetical cheating and justifications for hypothetical cheating through increasing the negative evaluations of the imagined learning situations. Thus, this work serves as first evidence for the linkage among tests as difficult learning tasks, resulting negative consequences like negative evaluations or stress perceptions, and hypothetical cheating.

Acknowledgements Open Access funding provided by Projekt DEAL. We thank Chawwah Grünberg, Sonja Haverland, Antonia Mariß, Luisa Neufeld, Celina Stolz, Tobias Steppat, and Laura Wagner for their help in conducting/programming the study. We thank Sarah Tyrrell for proofreading the manuscript.

Funding This research was supported by a LOEWE grant from the Hessian Ministry for Science and the Arts entitled "desirable difficulties; intrinsic cognitive motivation and performance expectancies" awarded to Marc-André Reinhard.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval The study was conducted in accordance with the Ethical Guidelines of the German Association of Psychologists (DGPs) as well as the American Psychological Association (APA), and the project was approved by the Ethics Committee affiliated with the funding source. Additionally, by the time the data were acquired it was neither customary at Kassel University nor at most other German universities to seek ethics approval for simple studies on personality, attitudes, and hypothetical decisions. Thus, ethical approval was not required for this study in accordance with the national and institutional guidelines. However, the study exclusively makes use of anonymous questionnaires and no identifying information was obtained from participants. Moreover, every participant had to read and agree to two questions concerning their consent. They were also explicitly informed that all their data would be treated confidentially and that they could withdraw from the study at any given time, without explanation, by simply closing the internet browser.

Appendix

Appendix A: Materials

Tests with public results learning scenario condition (including the instructions):

> This is a potential scenario that could happen in your daily life as a student. We would like to ask you to transport yourself in the situation, and to imagine it as strongly as you can. Imagine that you are a student in college and have lots of exams to write. During one of your majors your professor tries to increase your and your fellow students learning success, and enhance your chance to pass the exam. Therefore, half an hour before the end of every session you write an ungraded test, and answer multiple questions concerning the content of that session. Once the half an hour is up you can go home. Shortly following every session all students receive an e-mail with the matriculation numbers of everyone, and their test results, ranking from best to worst.

Tests with private results learning scenario condition (including the instructions):

> This is a potential scenario that could happen in your daily life as a student. We would like to ask you to transport yourself in the situation, and to imagine it as strongly as you can. Imagine that you are a student in college and have lots of exams to write. During one of your majors your professor tries to increase your and your fellow students learning success, and enhance your chance to pass the exam. Therefore, half an hour before the end of every session you write an ungraded test, and answer multiple questions concerning the content of that session. Once the half an hour is up you can go home. Shortly following every session you receive a private e-mail with your own test results.

Reading control learning scenario condition (including the instructions):

> This is a potential scenario that could happen in your daily life as a student. We would like to ask you to transport yourself in the situation, and to imagine it as strongly as you can. Imagine that you are a student in college and have lots of exams to write. During one of your majors your professor tries to increase your and your fellow students learning success and enhance your chance to pass the exam. Therefore, half an hour before the end of every session your professor hands you a summary with all the relevant information of that session. In this time you read the materials. Once the half an hour is up you can go home.

Negative Evaluations of the Learning Situations (10 items, $\alpha = .89$): Instructions and items

"Please answer the following questions according to your imagined mood/perception/thoughts/feelings during the situation displayed in the scenario."

"Concerning the imagined scenario, …"

1. How strenuous did you find the described and imagined learning-situation? One (*not strenuous at all*)—seven (*extremely strenuous*)
2. How (un)just did you find the described and imagined way of learning in the situation? One (*extremely unjust*)—seven (*extremely just*), *recoded item
3. How difficult would your rate the learning in the described situation? One (*not difficult at all*)—seven (*extremely difficult*)
4. How fair or unfair would you rate the way of learning in such a situation? One (*extremely unfair*)—seven (*extremely fair*), *recoded item
5. How angry would you be if you were in such a situation and had to learn in such a manner? One (*not in the least bit angry*)—seven (*extremely angry*)
6. How relaxing would you rate such a learning-situation? One (*not relaxing at all*)—seven (*extremely relaxing*), *recoded item
7. How overstrained would you feel if you were in such a learning-situation? One (*not at all*)—seven (*totally*)
8. How annoyed would you feel if you were in such a learning-situation? One (*not at all*)—seven (*extremely*)
9. How uncertain would you feel if you had to learn in a way as described in the situation? One (*not at all*)—seven (*extremely*)
10. How inappropriate would you rate such a learning-situation? One (*not inappropriate at all*)—seven (*extremely inappropriate*)

Not analysed positive control items:

1. How attentive would you be in such a learning-situation? One (*not attentive at all*)—seven (*extremely attentive*)
2. How interesting did you find the described learning-situation? One (*not a bit interesting*)—seven (*extremely interesting*)
3. How helpful and successful would you rate such a learning-situation? One (*not helpful or successful at all*)—seven (*extremely helpful and successful*)

Cheating items:

1. How likely is it that you would use your crib sheet to cheat in the exam? One (*Not at all likely, you would never use the crib sheet*)—seven (*Extremely likely, you would use the crib sheet definitely*)
2. How justifiable is it for you to use your crib sheet to cheat in the exam? One (*Not justifiable at all*)—seven (*Extremely justifiable*)
3. How likely is it that someone else would use their crib sheet to cheat in the exam if they were in the same situation? One (*Not at all likely, other people would never*

use the crib sheet)—seven (*Extremely likely, other people would use the crib sheet definitely*)
4. How justifiable is it for someone else to use their crib sheet to cheat in the exam? One (*Not justifiable at all*)—seven (*Extremely justifiable*)
5. Furthermore, how likely is it that you would intentionally prepare a cheat sheet with the intention to use it in the exam? One (*Not at all likely, you would never prepare a cheat sheet*)—seven (*Extremely likely, you would definitely prepare a cheat sheet*)
6. How justifiable is it for you to prepare a cheat sheet to cheat in the exam? One (*Not justifiable at all*)—seven (*Extremely justifiable*)
7. How likely is it that someone else would intentionally prepare a cheat sheet to use it in the exam if they were in the same situation? One (*Not at all likely, other people would never prepare a cheat sheet*)—seven (*Extremely likely, other people would definitely prepare a cheat sheet*)
8. How justifiable is it for someone else to prepare a cheat sheet to cheat in the exam? One (*Not justifiable at all*)—seven (*Extremely justifiable*)

Average across items 1 and 5: *likelihoods own cheating,* $\alpha = .86$
Average across items 3 and 7: *likelihoods others' cheating,* $\alpha = .84$
Average across items 2, 4, 6, and 8: *justifications for cheating,* $\alpha = .95$

General manipulation-check questions:

1. Have you really read and imagined the former scenario? *No, not at all/A little bit/Yes*
2. Did you understand the described scenario? One (*No, not at all*)—seven (*Yes, completely*)
3. Were you able to put yourself in the described scenario? One (*No, not at all*)—seven (*Yes, totally*)
4. Have you (in your daily life as a student) experienced situations similar to the ones described in the scenario? One (*No, never*)—seven (*Yes, multiple times*)

Manipulation-check questions concerning cheating:

"Given the former scenario, …"

1. How likeable would you rate your professor? One (*absolutely unlikeable*)—seven (*extremely likeable*)
2. How important are good grades for you? One (*absolutely unimportant*)—seven (*extremely important*)
3. How much do you think you can improve the results of your exam through cheating? One (*no improvement at all*)—seven (*extremely high improvement*)
4. How likely is it that you (if you decided to cheat) would get caught? One (*absolutely unlikely*)—seven (*extremely likely*)
5. How would you rate the consequences of cheating if you would get caught? One (*absolutely no consequences*)—seven (*extremely severe consequences*)

"In general, …"

6. How intense would you rate the pressure to perform during your study program? One (*not intense at all*)—seven (*very intense*)
7. Have you cheated in exams before? One (*no, never*)—seven (*yes, every time*)
8. Do you have negative or positive attitudes toward cheating during exams? One (*completely negative attitudes*)—seven (*completely positive attitudes*)

Appendix B: Further analyses

Factor analyses with Varimax rotation

We conducted two factor analyses to test the factor structure of the eight cheating items: Regarding the four likelihoods of cheating items, we conducted a factor analysis with varimax rotation. The eigen values and a scree plot yielded two factors explaining a total of 86.99% of the variance of all four likelihoods items (see Table 3). Factor 1 consisted of the two items regarding the likelihoods of own cheating (*likelihoods own cheating,* $\alpha = .86$; 58.98% of explained variance, eigen value = 2.36). Factor 2 consisted of the two items regarding the likelihoods of others' cheating (*likelihoods others' cheating,* $\alpha = .84$; 28.01% of explained variance, eigen value = 1.12). The second factor analysis with varimax rotation was conducted with the four justifications for cheating items. The eigen values and a scree plot yielded one factor explaining a total of 86.23% of the variance of all four items (*justifications for cheating,* $\alpha = .95$; eigen value = 3.45; see Table 4).

Table 3 Factor analysis likelihoods of hypothetical cheating variables

Variables	Factor loading	
	1	2
Likelihoods own spontaneous cheating	.920	.179
Likelihoods others' spontaneous cheating	.171	.912
Likelihoods own prepared cheating	.924	.161
Likelihoods others' prepared cheating	.164	.913

Table 4 Factor analysis justifications for cheating variables

Variables	Factor loading
	1
Justifications own spontaneous cheating	.922
Justifications others' spontaneous cheating	.938
Justifications own prepared cheating	.936
Justifications others' prepared cheating	.919

Table 5 Correlations among the negative evaluations of the learning situations, stress perceptions (PSQ), likelihoods own cheating, likelihoods others' cheating, justifications for cheating, trait test anxiety (PAF-E), trait stress scales (PSS, SSS), and academic self-concept

	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Negative evaluation of the learning situation	1								
2. Stress perceptions (PSQ)	.50**	1							
3. Likelihoods own cheating	.19**	.08^{+}	1						
4. Likelihoods others' cheating	.04	.14**	.36**	1					
5. Justifications for cheating	.16**	.08	.86**	.33**	1				
6. Trait test anxiety (PAF-E)	.28**	.56**	.14**	.14**	.11*	1			
7. Trait stress (PSS)	.33**	.52**	.11*	.10^{+}	.07	.65**	1		
8. Trait stress (SSS)	.31**	.54**	.08	.07	.05	.60**	.67**	1	
9. Academic self-concept	− .18**	− .27**	− .22**	− .05	− .25**	− .43**	− .40**	− .29**	1

$^{+}p < .10$, $*p < .05$, $**p < .01$, Two-tailed. $N = 405$

We further conducted a MANCOVA predicting participants trait test anxiety (PAF-E), stress traits (PSS, SSS), and their academic self-concept using the learning scenario condition as the between-subjects variable to test for differences between participants in the three learning scenario conditions. The one-way MANCOVA yielded no significant multivariate main effect for the learning scenario condition, $F(8,800) = .50$, $p = .859$, $\eta_p = .005$. Given the not significant overall test, the univariate main effects were not examined. We also conducted correlations among the negative evaluations of the learning situations, stress perceptions (PSQ), likelihoods own cheating, likelihoods others' cheating, and justifications for cheating with the trait variables (trait test anxiety, trait stress, and academic self-concept) that were solely assessed for the other study that was conducted together with this study (see Table 5).

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica, 87,* 1115–1153. https://doi.org/10.3982/ECTA14673.

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87,* 659–701. https://doi.org/10.3102/0034654316689306.

Agnew, R. (1992). Foundation for a general strain theory of crime and delinquency. *Criminology, 30,* 47–88. https://doi.org/10.1111/j.1745-9125.1992.tb01093.x.

Agnew, R., & White, H. R. (1992). An empirical test of general strain theory. *Criminology, 30*(4), 475–500. https://doi.org/10.1111/j.1745-9125.1992.tb01113.x.

Akers, R. L. (1990). Rational choice, deterrence, and social learning theory in criminology: The path not taken. *Journal of Criminal Law & Criminology, 81,* 653–676. https://doi.org/10.2307/1143850.

Alter, A. L., Oppenheimer, D. M., & Epley, N. (2013). Disfluency prompts analytic thinking—But not always greater accuracy: Response to Thompson et al. (2013). *Cognition, 128,* 252–255. https://doi.org/10.1016/j.cognition.2013.01.006.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136,* 569. https://doi.org/10.1037/0096-3445.136.4.569.

Anderman, E. M., Griesinger, T., & Westerfield, G. (1998). Motivation and cheating during early adolescence. *Journal of Educational Psychology, 90,* 84–93. https://doi.org/10.1037/0022-0663.90.1.84.

Becker, G. S. (1968). Crime and punishment: An economic approach. In G. S. Becker & W. M. Landes (Eds.), *The economic dimensions of crime* (pp. 13–68). London: Palgrave Macmillan.

Bertrams, A., & Englert, C. (2014). Test anxiety, self-control, and knowledge retrieval in secondary school students. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 46,* 165–170. https://doi.org/10.1026/0049-8637/a000111.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35,* 201–210. https://doi.org/10.3758/BF03193441.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge: MIT Press.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge: The MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Boca Raton: CRC Press.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society, 2,* 59–68.

Brimble, M., & Stevenson-Clarke, P. (2005). Perceptions of the prevalence and seriousness of academic dishonesty in Australian universities. *The Australian Educational Researcher, 32,* 19–44. https://doi.org/10.1007/BF0321682.

Calabrese, R. L., & Cochran, J. T. (1990). The relationship of alienation to cheating among a sample of American adolescents. *Journal of Research & Development in Education, 23*(2), 65–72.

Cameron, J. S., Miller, D., & Monin, B. (2008). Deservingness and unethical behavior in loss and gain frames. *Unpublished manuscript*. http://www.communicationcache.com/uploads/1/0/8/8/10887248/deservingness_and_unethical_behavior_in_loss_gain_frames.pdf.

Campbell, W. K., Bonacci, A. M., Shelton, J., Exline, J. J., & Bushman, B. J. (2004). Psychological entitlement: Interpersonal consequences and validation of a self-report measure. *Journal of Personality Assessment, 83,* 29–45. https://doi.org/10.1207/s15327752jpa8301_04.

Carmichael, S., & Piquero, A. R. (2004). Sanctions, perceived anger, and criminal offending. *Journal of Quantitative Criminology, 20,* 371–393. https://doi.org/10.1007/s10940-004-5869-y.

Carrell, S. E., Malmstrom, F. V., & West, J. E. (2008). Peer effects in academic cheating. *Journal of Human Resources, 43,* 173–207. https://doi.org/10.3368/jhr.43.1.173.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132,* 354–380. https://doi.org/10.1037/0033-2909.132.3.354.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*. https://doi.org/10.2307/2136404.

Davis, S. F., Grover, C. A., Becker, A. H., & McGregor, L. N. (1992). Academic dishonesty: Prevalence, determinants, techniques, and punishments. *Teaching of Psychology, 19,* 16–20. https://doi.org/10.1207/s15328023top1901_3.

De Bruin, G. P., & Rudnick, H. (2007). Examining the cheats: The role of conscientiousness and excitement seeking in academic dishonesty. *South African Journal of Psychology, 37,* 153–164. https://doi.org/10.1177/008124630703700111.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology, 70,* 979–995. https://doi.org/10.1037/0022-3514.70.5.979.

Dickhäuser, O., Schöne, C., Spinath, B., & Stiensmeier-Pelster, J. (2002). Die Skalen zum akademischen Selbstkonzept: Konstruktion und Überprüfung eines neuen Instrumentes [The academic self-concept scales: Construction and evaluation of a new instrument]. *Zeitschrift für differentielle und diagnostische Psychologie: ZDDP, 23,* 393–405. https://doi.org/10.1024//0170-1789.23.4.393.

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the bold (and the italicized): Effects of disfluency on educational outcomes. *Cognition, 118,* 111–115. https://doi.org/10.1016/j.cognition.2010.09.012.

Dobson, J. L., & Linderholm, T. (2015). The effect of selected "desirable difficulties" on the ability to recall anatomy information. *Anatomical Sciences Education, 8,* 395–403. https://doi.org/10.1002/ase.1489.

Doménech-Betoret, F., Abellán-Roselló, L., & Gómez-Artiga, A. (2017). Self-efficacy, satisfaction, and academic achievement: The mediator role of Students' expectancy-value beliefs. *Frontiers in Psychology, 8,* 1193. https://doi.org/10.3389/fpsyg.2017.01193.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4–58. https://doi.org/10.1177/1529100612453266.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,* 1149–1160. https://doi.org/10.3758/BRM.41.4.1149.

Feldman, R. S., Forrest, J. A., & Happ, B. R. (2002). Self-presentation and verbal deception: Do self-presenters lie more? *Basic and Applied Social Psychology, 24,* 163–170. https://doi.org/10.1207/S15324834BASP2402_8.

Fida, R., Paciello, M., Tramontano, C., Fontaine, R. G., Barbaranelli, C., & Farnese, M. L. (2015). An integrative approach to understanding counterproductive work behavior: The roles of stressors, negative emotions, and moral disengagement. *Journal of Business Ethics, 130,* 131–144. https://doi.org/10.1007/s10551-014-2209-5.

Fida, R., Tramontano, C., Paciello, M., Ghezzi, V., & Barbaranelli, C. (2018). Understanding the interplay among regulatory self-efficacy, moral disengagement, and academic cheating behaviour during vocational education: A three-wave study. *Journal of Business Ethics, 153,* 725–740. https://doi.org/10.1007/s10551-016-3373-6.

Finn, K. V., & Frone, M. R. (2004). Academic performance and cheating: Moderating role of school identification and self-efficacy. *The Journal of Educational Research, 97,* 115–121. https://doi.org/10.3200/JOER.97.3.115-121.

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise—An experimental study on cheating. *Journal of the European Economic Association, 11,* 525–547. https://doi.org/10.1111/jeea.12014.

Franklyn-Stokes, A., & Newstead, S. E. (1995). Undergraduate cheating: Who does what and why? *Studies in Higher Education, 20,* 159–172. https://doi.org/10.1080/03075079512331381673.

Freiburger, T. L., Romain, D. M., Randol, B. M., & Marcum, C. D. (2017). Cheating behaviors among undergraduate college students: Results from a factorial survey. *Journal of Criminal Justice Education, 28,* 222–247. https://doi.org/10.1080/10511253.2016.1203010.

Geißler, H., Schöpe, S., Klewes, J., Rauh, C., & von Alemann, U. (2013). *Wertestudie 2013: Wie groß ist die Kluft zwischen dem Volk und seinen Vertretern [Value Study 2013: How big is the gap between the people and their representatives].* Köln: YouGov.

Giluk, T. L., & Postlethwaite, B. E. (2015). Big five personality and academic dishonesty: A meta-analytic review. *Personality and Individual Differences, 72,* 59–67. https://doi.org/10.1016/j.paid.2014.08.027.

Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological Science, 20,* 393–398. https://doi.org/10.1111/j.1467-9280.2009.02306.x.

Greenberg, J. (1990). Employee theft as a reaction to underpayment inequity: The hidden cost of pay cuts. *Journal of Applied Psychology, 75,* 561–568. https://doi.org/10.1037/0021-9010.75.5.561.

Greene, A. S., & Saxe, L. (1992). *Everybody (else) does it: Academic cheating* [Paper presentation]. Annual meeting of the Eastern Psychological Association, April 3–5, Boston, MA.

Haines, V. J., Diekhoff, G. M., LaBeff, E. E., & Clark, R. E. (1986). College cheating: Immaturity, lack of commitment, and the neutralizing attitude. *Research in Higher Education, 25,* 342–354. https://doi.org/10.1007/BF00992130.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis* (2nd ed.). New York: The Guilford Press.

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: Performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology, 28,* 597–606. https://doi.org/10.1002/acp.3032.

Hoferichter, F., Raufelder, D., Ringeisen, T., Rohrmann, S., & Bukowski, W. M. (2016). Assessing the multi-faceted nature of test anxiety among secondary school students: An English version of the German test anxiety questionnaire: PAF-E. *The Journal of Psychology, 150,* 450–468. https://doi.org/10.1080/00223980.2015.1087374.

Hoffman, E., & Spitzer, M. L. (1985). Entitlements, rights, and fairness: An experimental examination of subjects' concepts of distributive justice. *The Journal of Legal Studies, 14*(2), 259–297.

Houser, D., Vetter, S., & Winter, J. (2012). Fairness and cheating. *European Economic Review, 56,* 1645–1655. https://doi.org/10.1016/j.euroecorev.2012.08.001.

Ikeda, K., Castel, A. D., & Murayama, K. (2015). Mastery-approach goals eliminate retrieval-induced forgetting: The role of achievement goals in memory inhibition. *Personality and Social Psychology Bulletin, 41,* 687–695. https://doi.org/10.1177/0146167215575730.

Jensen, L. A., Arnett, J. J., Feldman, S. S., & Cauffman, E. (2002). It's wrong, but everybody does it: Academic dishonesty among high school and college students. *Contemporary Educational Psychology, 27,* 209–228. https://doi.org/10.1006/ceps.2001.1088.

Kaiser, I., Mayer, J., & Malai, D. (2018). Self-generation in the context of inquiry-based learning. *Frontiers in Psychology, 9,* 2440. https://doi.org/10.3389/fpsyg.2018.02440.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93,* 579–588. https://doi.org/10.1037/0022-0663.93.3.579.

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17,* 471–479. https://doi.org/10.1080/09658210802647009.

Karpicke, J. D., & Roediger, H. L., III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33,* 704–719. https://doi.org/10.1037/0278-7393.33.4.704.

Kausar, R. (2010). Perceived stress, academic workloads and use of coping strategies by university students. *Journal of Behavioural Sciences, 20*(1), 31–45.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22*(6), 787–794. https://doi.org/10.1177/0956797611407929.

Koul, R. (2012). Multiple motivational goals, values, and willingness to cheat. *International Journal of Educational Research, 56,* 1–9. https://doi.org/10.1016/j.ijer.2012.10.002.

LaBeff, E. E., Clark, R. E., Haines, V. J., & Diekhoff, G. M. (1990). Situational ethics and college student cheating. *Sociological Inquiry, 60*(2), 190–198. https://doi.org/10.1111/j.1475-682X.1990.tb00138.x.

Lehmann, J., Goussios, C., & Seufert, T. (2016). Working memory capacity and disfluency effect: An aptitude-treatment-interaction study. *Metacognition and Learning, 11,* 89–105. https://doi.org/10.1007/s11409-015-9149-z.

Leiner, J. E. M., Scherndl, T., & Ortner, T. M. (2018). How do men and women perceive a high-stakes test situation? *Frontiers in Psychology, 9,* 2216. https://doi.org/10.3389/fpsyg.2018.02216.

Levenstein, S., Prantera, C., Varvo, V., Scribano, M. L., Berto, E., Luzi, C., et al. (1993). Development of the perceived stress questionnaire: A new tool for psychosomatic research. *Journal of Psychosomatic Research, 37,* 19–32. https://doi.org/10.1016/0022-3999(93)90120-5.

Lipowsky, F., Richter, T., Borromeo-Ferri, R., Ebersbach, M., & Hänze, M. (2015). Wünschenswerte Erschwernisse beim Lernen [Desirable difficulties during learning]. *Schulpädagogik heute, 6*(11), 1–10.

Marcela, V. (2015). Learning strategy, personality traits and academic achievement of university students. *Procedia-Social and Behavioral Sciences, 174,* 3473–3478. https://doi.org/10.1016/j.sbspro.2015.01.1021.

Marksteiner, T., Reinhard, M. A., Lettau, F., & Dickhäuser, O. (2013). Bullying, cheating, deceiving: Teachers' perception of deceitful situations at school. *International Journal of Educational Psychology, 2,* 193–220. https://doi.org/10.4471/ijep.2013.24.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research, 45,* 633–644. https://doi.org/10.1509/jmkr.45.6.633.

McCabe, D. L. (1992). The influence of situational ethics on cheating among college students. *Sociological Inquiry, 62,* 365–374. https://doi.org/10.1111/j.1475-682X.1992.tb00287.x.

McCabe, D. L. (2001). *Academic integrity—A research update*. College Station: Center for Academic Integrity.

McCabe, D. L., Treviño, L. K., & Butterfield, K. D. (2001). Cheating in academic institutions: A decade of research. *Ethics and Behavior, 11,* 219–232. https://doi.org/10.1207/S15327019EB1103_2.

McDaniel, M. A., Hines, R. J., & Guynn, M. J. (2002). When text difficulty benefits less-skilled readers. *Journal of Memory and Language, 46,* 544–561. https://doi.org/10.1006/jmla.2001.2819.

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14,* 200–206. https://doi.org/10.3758/BF03194052.

McDaniel, M. A., Waddill, P. J., & Einstein, G. O. (1988). A contextual account of the generation effect: A three-factor theory. *Journal of Memory and Language, 27,* 521–536. https://doi.org/10.1016/0749-596X(88)90023-X.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14,* 1–43. https://doi.org/10.1207/s1532690xci1401_1.

Messick, D. M., Bloom, S., Boldizar, J. P., & Samuelson, C. D. (1985). Why we are fairer than others. *Journal of Experimental Social Psychology, 21,* 480–500. https://doi.org/10.1016/0022-1031(85)90031-9.

Mihalca, L., Mengelkamp, C., & Schnotz, W. (2017). Accuracy of metacognitive judgments as a moderator of learner control effectiveness in problem-solving tasks. *Metacognition and Learning, 12,* 357–379. https://doi.org/10.1007/s11409-017-9173-2.

Naghdipour, B., & Emeagwali, O. L. (2013). Students' justifications for academic dishonesty: Call for action. *Procedia-Social and Behavioral Sciences, 83,* 261–265. https://doi.org/10.1016/j.sbspro.2013.06.051.

Newstead, S. E., Franklyn-Stokes, A., & Armstead, P. (1996). Individual differences in student cheating. *Journal of Educational Psychology, 88,* 229–241. https://doi.org/10.1037/0022-0663.88.2.229.

Nonis, S., & Swift, C. O. (2001). An examination of the relationship between academic dishonesty and workplace dishonesty: A multicampus investigation. *Journal of Education for Business, 77,* 69–77. https://doi.org/10.1080/08832320109599052.

Olafson, L., Schraw, G., Nadelson, L., Nadelson, S., & Kehrwald, N. (2013). Exploring the judgment–action gap: College students and academic dishonesty. *Ethics and Behavior, 23,* 148–162. https://doi.org/10.1080/10508422.2012.714247.

O'Neil, J. H., Spielberger, C. D., & Hansen, D. N. (1969). Effects of state anxiety and task difficulty on computer-assisted learning. *Journal of Educational Psychology, 60,* 343–350. https://doi.org/10.1037/h0028323.

Oppenheimer, D. M., & Alter, A. L. (2014). The search for moderators in disfluency research. *Applied Cognitive Psychology, 28,* 502–504. https://doi.org/10.1002/acp.3023.

Paternoster, R., McGloin, J. M., Nguyen, H., & Thomas, K. J. (2013). The causal impact of exposure to deviant peers: An experimental investigation. *Journal of Research in Crime and Delinquency, 50,* 476–503. https://doi.org/10.1177/0022427812444274.

Paulhus, D. L., & Dubois, P. J. (2015). The link between cognitive ability and scholastic cheating: A meta-analysis. *Review of General Psychology, 19,* 183–190. https://doi.org/10.1037/gpr0000040.

Pieger, E., Mengelkamp, C., & Bannert, M. (2016). Metacognitive judgments and disfluency—Does disfluency lead to more accurate judgments, better control, and better performance? *Learning and Instruction, 44,* 31–40. https://doi.org/10.1016/j.learninstruc.2016.01.012.

Reeder, L. G., Schrama, P. G. M., & Dirken, J. M. (1973). Stress and cardiovascular health: An international cooperative study—I. *Social Science and Medicine, 1967*(7), 573–584. https://doi.org/10.1016/0037-7856(73)90026-7.

Reinhard, M. A., Dickhäuser, O., Marksteiner, T., & Sporer, S. L. (2011). The case of Pinocchio: Teachers' ability to detect deception. *Social Psychology of Education, 14,* 299–318. https://doi.org/10.1007/s11218-010-9148-5.

Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). Linking cognitive science to education: Generation and interleaving effects. In *Proceedings of the twenty-seventh annual conference of the Cognitive Science Society* (pp. 1850–1855). Mahwah: Erlbaum.

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x.

Roelle, J., & Berthold, K. (2017). Effects of incorporating retrieval into learning tasks: The complexity of the tasks matters. *Learning and Instruction, 49,* 142–156. https://doi.org/10.1016/j.learninstruc.2017.01.008.

Rost, D. H., & Wild, K. P. (1994). Cheating and achievement-avoidance at school: Components and assessment. *British Journal of Educational Psychology, 64,* 119–132. https://doi.org/10.1111/j.2044-8279.1994.tb01089.x.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*. https://doi.org/10.1037/a0037559.

Sarason, I. G., & Sarason, B. R. (1990). Test anxiety. *Handbook of social and evaluation anxiety* (pp. 475–495). Boston: Springer. https://doi.org/10.1007/978-1-4899-2504-6_16.

Schab, F. (1991). Schooling without learning: Thirty years of cheating in high school. *Adolescence, 26*(104), 839–847.

Schunk, D. H. (1996). *Self-efficacy for learning and performance*. [Paper presentation]. American Educational Research Association, April 8–12, New York, NY.

Schunk, D. H., & Gaa, J. P. (1981). Goal-setting influence on learning and self-evaluation. *The Journal of Classroom Interaction, 16*(2), 38–44.

Shalvi, S., Dana, J., Handgraaf, M. J., & De Dreu, C. K. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes, 115,* 181–190. https://doi.org/10.1016/j.obhdp.2011.02.001.

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-serving justifications: Doing wrong and feeling moral. *Current Directions in Psychological Science, 24,* 125–130. https://doi.org/10.1177/0963721414553264.

Shu, L. L., Gino, F., & Bazerman, M. H. (2011). Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality and Social Psychology Bulletin, 37,* 330–349. https://doi.org/10.1177/0146167211398138.

Simha, A., & Cullen, J. B. (2012). A comprehensive literature review on cheating. *International Journal of Cyber Ethics in Education (IJCEE), 2,* 24–44. https://doi.org/10.4018/ijcee.2012100102.

Steininger, M., Johnson, R. E., & Kirts, D. K. (1964). Cheating on college examinations as a function of situationally aroused anxiety and hostility. *Journal of Educational Psychology, 55,* 317–324. https://doi.org/10.1037/h0042396.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction, 12,* 185–233. https://doi.org/10.1207/s1532690xci1203_1.

Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 607. https://doi.org/10.1037/0278-7393.5.6.607.

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: the testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review, 27,* 247–264. https://doi.org/10.1007/s10648-015-9310-x.

Wenzel, K., & Reinhard, M. A. (2019a). Relatively unintelligent individuals do not benefit from intentionally hindered learning: The role of desirable difficulties. *Intelligence, 77,* 101405. https://doi.org/10.1016/j.intell.2019.101405.

Wenzel, K., & Reinhard, M. A. (2019b). Does the end justify the means? Learning tests lead to more negative evaluations and to more stress experiences. *Manuscript submitted for publication*.

Whitley, B. E. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education, 39,* 235–274. https://doi.org/10.1023/A:1018724900565.

Wowra, S. A. (2007). Moral identities, social anxiety, and academic dishonesty among American college students. *Ethics and Behavior, 17,* 303–321. https://doi.org/10.1080/10508420701519312.

Yu, H., Glanzer, P. L., Sriram, R., Johnson, B. R., & Moore, B. (2017). What contributes to college students' cheating? A study of individual factors. *Ethics and Behavior, 27,* 401–422. https://doi.org/10.1080/10508422.2016.1169535.

Kristin Wenzel is a Ph.D. student of Social Psychology at the Department of Psychology, University of Kassel, Germany. Her research interests include cognitive prerequisites and moderators of the positive effect of desirable difficulties as well as consequences of desirable difficulties. Additionally, her research focuses on deception and deception detection as well as on the belief in a just world.

Marc-André Reinhard is a Full Professor of Social Psychology at the Department of Psychology, University of Kassel, Germany. His research interests include cognitive and motivational moderators of the positive effect of desirable difficulties in learning. Moreover, his research focuses on the formation and consequences of performance expectancies.

APPENDIX F

Wenzel, K., & Reinhard, M.-A. (2021b). Learning with a double-edged sword? Beneficial and detrimental effects of learning tests—taking a first look at linkages among tests, later learning outcomes, stress perceptions, and intelligence. *Frontiers in Psychology*, *12*, Article 693585. https://doi.org/10.3389/fpsyg.2021.693585

This is the final article version published by Frontiers in *Frontiers in Psychology* available online: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.693585/full

frontiers in Psychology

ORIGINAL RESEARCH
published: 31 August 2021
doi: 10.3389/fpsyg.2021.693585

Learning With a Double-Edged Sword? Beneficial and Detrimental Effects of Learning Tests—Taking a First Look at Linkages Among Tests, Later Learning Outcomes, Stress Perceptions, and Intelligence

Kristin Wenzel and Marc-André Reinhard*

Department of Psychology, University of Kassel, Kassel, Germany

OPEN ACCESS

Edited by:
Lu Wang,
University of Georgia,
United States

Reviewed by:
Petar Radanliev,
University of Oxford,
United Kingdom
Jamie J. Jirout,
University of Virginia,
United States

****Correspondence:***
Kristin Wenzel
kristin.wenzel@uni-kassel.de

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *11 April 2021*
Accepted: *28 July 2021*
Published: *31 August 2021*

Citation:
Wenzel K and Reinhard M-A (2021) Learning With a Double-Edged Sword? Beneficial and Detrimental Effects of Learning Tests—Taking a First Look at Linkages Among Tests, Later Learning Outcomes, Stress Perceptions, and Intelligence. Front. Psychol. 12:693585. doi: 10.3389/fpsyg.2021.693585

It has often been shown that tests as intentionally hindered and difficult learning tasks increase long-term learning compared to easier tasks. Previous work additionally indicated that higher intelligence might serve as a prerequisite for such beneficial effects of tests. Nevertheless, despite their long-term learning effects, tests were also found to be evaluated as more negative and to lead to more stress and anxiety compared to easier control tasks. Stress and anxiety, in turn, often yield detrimental effects on learning outcomes. Hence, we hypothesized that tests increase later learning outcomes but simultaneously also lead to more stress perceptions. Such increased stress was, in turn, hypothesized to reduce later learning outcomes (thus, stress might serve as a mediator of the beneficial effects of tests on learning). All these assumed effects should further be moderated by intelligence, insofar as that higher intelligence should increase beneficial effects of tests on learning, should decrease stress perceptions caused by tests, and should reduce detrimental effects of stress on learning outcomes. Higher intelligence was also assumed to be generally associated with higher learning. We conducted a laboratory study ($N = 89$) to test these hypotheses: Participants underwent an intelligence screening, then worked on either a test or a re-reading control task, and reported their immediate stress perceptions. Later learning outcomes were assessed after 1 week. The results supported all assumed main effects but none of the assumed interactions. Thus, participants using tests had higher long-term learning outcomes compared to participants using re-reading tasks. However, participants using tests also perceived more immediate stress compared to participants that only re-read the materials. These stress perceptions in turn diminished the beneficial effects of tests. Stress was also generally related to lower learning, whereas higher intelligence was linked to higher learning and also to lower stress. Hence, our findings again support the often assumed benefits of tests—even when simultaneously considering learners' intelligence and and when considering the by tests caused stress perceptions.

Notably, controlling for stress further increases these long-term learning benefits. We then discuss some limitations and boundaries of our work as well as ideas for future studies.

Keywords: learning tests, desirable difficulties, acute stress perceptions, intelligence, long-term learning

INTRODUCTION

The following work raises the question if normally beneficial learning tests actually serve as double-edged swords, thus, if they can result in both beneficial as well as detrimental effects: More specifically, the present work was conducted to simultaneously focus on the often observed positive long-term learning effects of tests as difficult and demanding learning strategies (see, e.g., Adesope et al., 2017; Yang et al., 2021) but also on potential negative (side) effects caused by such learning tests, namely, increased stress or anxiety perceptions (see, e.g., Hinze and Rapp, 2014; Wenzel and Reinhard, 2021). Such increased stress perceptions should have further detrimental effects on learning in general as well as on the beneficial effects of tests on long-term learning in specific (see, e.g., Seipp, 1991; Hinze and Rapp, 2014). Additionally, because recent studies indicated that higher intelligence is valuable for the effectiveness of tests (see, e.g., Minear et al., 2018; Wenzel and Reinhard, 2019), the present work also investigates if higher intelligence moderates the benefits of tests, thus serving as a prerequisite or boundary condition. In line with this, different previous studies indirectly supported the assumption that intelligence might also act as a buffer for negative effects of tests on immediate stress perceptions (see, e.g., LePine et al., 2004; Abín et al., 2020) and for the detrimental effects of stress perceptions on learning outcomes (see, e.g., Chuderski, 2014; Reeve et al., 2014). Hence, the present work bridges different research fields and simultaneously focuses on beneficial and detrimental effects of tests as well as on potentially moderating effects of intelligence as an important individual difference. Simultaneously testing these different research issues seems necessary for being able to give empirically well-grounded advice regarding the application of tests in university or school settings to learners and lecturers alike—especially because we not only investigate learning outcomes but also students' experiences and perceptions as well as individual differences as potential prerequisites.

More specifically, focusing on these research questions is extremely relevant due to the importance of successful and durable later learning outcomes in school and university settings. Notably, although difficult learning strategies, like tests, have often been shown to increase long-term learning compared to learning strategies that are more fluent and simpler, learners and lecturers mainly assume the contrary (e.g., Karpicke et al., 2009; Diemand-Yauman et al., 2011; Kornell et al., 2011; Dobson and Linderholm, 2015; Bjork and Bjork, 2019). Thus, learners normally regard easy and fluent learning strategies as more effective and most prefer simpler strategies, like repeated reading—and such misconceptions even stick with teachers-to-be (e.g., Book et al., 1983; Koriat and Ma'ayan, 2005; Karpicke et al., 2009; Bjork et al., 2015). Hence, it is extremely important to conduct further empirical work to be able to give well-grounded advice to learners and lecturers alike that—or if—difficult tests are helpful and should be applied in actual university learning settings. Otherwise, they might not apply such tasks on their own. In line with this, lecturers and teachers often express concerns about the effectiveness of such difficult learning strategies for all of their students (e.g., Diemand-Yauman et al., 2011; Lipowsky et al., 2015), which is why we also test the importance of (higher) intelligence as a prerequisite for the beneficial effects of tests. This is relevant as it could further specify for which group of learners tests are beneficial and for which they are not. We thereby choose intelligence as an individual difference because it was often cited as one of the strongest predictors for academic achievement and is generally strongly associated with varying operationalizations of successful human behavior (see, e.g., Bornstein et al., 2013; Strenze, 2015). Surprisingly, we could not find much research concerning potential moderating effects of intelligence on the effectiveness of tests for long-term learning outcomes. In addition, and apart from such later learning outcomes, we also focus on learners' perceptions of tests to explore if these normally beneficial learning tasks also lead to negative side-effects like increased immediate stress perceptions during and directly after learning. This seems relevant because it is often argued that students' experiences and perceptions of different situations are seldom the main focus of experiments (see, e.g., Edwards and Templeton, 2005)—even though stress perceptions include, among others, subjective distress, higher degrees of worry, emotionality, tension, anxiety, nervousness, pressure, intrusive and disturbing thoughts, feelings of overwhelm, and lack of confidence (see, e.g., Epel et al., 2018). Hence, such stress perceptions in themselves are extremely unpleasant and undesirable but were additionally often shown to lead to further negative consequences like reduced motivation, mood disturbances, or health problems (e.g., DeLongis et al., 1988; Hobfoll, 1989; LePine et al., 2004). In line with this, stress perceptions have often been shown to be associated with lower learning outcomes (e.g., Seipp, 1991), so that stress perceptions might even act as a mediator of the beneficial effects of tests on later learning outcomes. Notably, this would be completely inconsistent with the intention of using tests in schools or universities and should therefore be thoroughly explored. Thus, it is extremely important to know if tests—even those conducted as low-stakes learning situations—lead to negative consequences, like increased stress perceptions, and if these would, paradoxically, be linked to reduced benefits of tests. It is also important to determine whether these negative side effects of tests on stress perceptions and the detrimental effects of stress on later learning outcomes arise for all learners or only for those with lower cognitive abilities. Hence, we also test if intelligence moderates these effects, thus, if immediate stress

perceptions caused by tests or detrimental effects of stress perceptions on later learning outcomes decrease with higher intelligence. This would indicate that intelligence might also serve as a protective factor for potentially negative side effects caused by such learning tests and for detrimental effects of acute stress perceptions. In turn, such findings might further help to specify for whom tests are actually desirable. Taken together, focusing on and answering these research questions is very important regarding potential advice for teachers and lecturers concerning the utilization and practical application of learning tests in schools and universities. We further think that the present work focuses on new and extremely relevant issues while also trying to replicate previous findings (e.g., the benefits of tests as well as increased stress perceptions due to tests) that are of great relevance for the research field. Moreover, to our knowledge, no previous studies were conducted to test these assumptions, and none simultaneously tested prerequisites, beneficial effects, and potentially detrimental effects of tests. Hence, we want to highlight these important issues and stimulate future research. In the following, we want to start with presenting a state of the art literature overview regarding our posed research issues.

Tests As Desirable Difficulties for Learning

Due to the importance of learning, knowledge acquisition, and academic achievement, a lot of researchers investigated varying learning strategies that improve durable long-term learning: For instance, *desirable difficulties* as challenging, demanding, and non-fluent learning processes have often been found to enhance later long-term learning outcomes compared to easier and more fluent learning processes (e.g., Bjork, 1994; Karpicke et al., 2009; Bjork and Bjork, 2011, 2020). Thus, although these effortful learning strategies appear to slow the learning process down at first and cause difficulties and challenges for learners, they increase information processing, retrieval, transfer, and ultimately leaners long-term learning (e.g., Bjork and Bjork, 2011, 2019, 2020). The term desirable difficulties thereby acts as an umbrella term for different intentionally hindered learning strategies, which lead to beneficial effects for later long-term learning outcomes: These include, for instance, *disfluency* (using harder-to-read fonts; Diemand-Yauman et al., 2011) and *generation* (generating materials and solutions instead of passive consumption; Bertsch et al., 2007). One especially robust desirable difficulty is the application of tests (also: *testing, testing effect, retrieval practice, test-enhanced learning, and learning/practice tests*): Taking (learning) tests on previously studied materials increases long-term learning compared to easier and more passive re-reading tasks or compared to note-taking as a stronger control task—even concerning a multitude of difficult, complex, and curricular subjects in realistic learning contexts (e.g., McDaniel et al., 2007; Dunlosky et al., 2013; Rowland, 2014; Karpicke and Aue, 2015; Adesope et al., 2017; Batsell et al., 2017; Rummer et al., 2017; Yang et al., 2021). These beneficial effects of tests were, among others, found for different types of learning materials (e.g., factual information, vocabulary, conceptual information, longer scientific textbook paragraphs, traditional (live) lectures/lessons, and recorded e-lectures/video-presentations) and for different types of test questions (e.g., multiple-choice questions, short-answer questions, fill-in-the-blank questions, comprehension-based questions, application-based questions, transfer questions, and inferences; e.g., Roediger and Karpicke, 2006; McDaniel et al., 2011, 2013; Dunlosky et al., 2013; Rowland, 2014; Khanna, 2015; Jing et al., 2016; Adesope et al., 2017; Iwamoto et al., 2017; Heitmann et al., 2018; Feraco et al., 2020; Yang et al., 2021). Moreover, tests were beneficial in varying (face-to-face or online) settings (e.g., laboratories, universities, classrooms, and at home/outside of class) and for students of different age groups (e.g., elementary school students, high school students, and university students; e.g., McDaniel et al., 2007, 2011; Roediger et al., 2011; Rowland, 2014; Adesope et al., 2017; Yang et al., 2021). Notably, the benefits of tests were also shown to arise when tests were administered in varying (conventional, computerized, or technological) modalities (e.g., paper-pencil tests, orally delivered tests, tests administered with computers, tests administered on online-websites, tests using clicker response systems, tests applied with mobile devices, and tests conducted with online applications like Kahoot; see, e.g., McDaniel et al., (2013), Grimaldi and Karpicke, (2014), Feraco et al., (2020), Wang and Tahir, (2020), Yang et al., (2021). Thus, researchers often recommend the application of tests as an effective learning task to increase learners long-term learning outcomes.

Theoretically, these beneficial effects of tests are often attributed to the stimulation of cognitive processes that increase the understanding, deeper semantic/cognitive processing, and encoding of information (e.g., Bjork, 1994; Bjork and Bjork, 2011; Dunlosky et al., 2013; Rowland, 2014). Tests are also supposed to lead to more analytic and elaborative thinking, more (effortful) retrieval practice, better anchoring of the learned information in long-term memory, and to an allocation of more effort and more cognitive resources while learning (e.g., Bjork and Bjork, 1992, 2011; Dunlosky et al., 2013; Rowland, 2014). Most important, the beneficial effects of tests are often argued to be stronger when the applied tests are more difficult and thereby elicit more difficult retrieval practice, when the test questions increase the depth of the required retrieval, and when learners have to indulge in more effort to work on and to solve the test questions (e.g., Tyler et al., 1979; Alter et al., 2007; Pyc and Rawson, 2009; Rowland, 2014; Maass and Pavlik, 2016; Greving and Richter, 2018). Tests were also shown to be more beneficial the more information learners were able to successfully retrieve and the more test questions they could answer correctly (e.g., Richland et al., 2005; Rowland, 2014). In line with this, previous work also yielded that desirable difficulties only increase long-term learning for learners who possess sufficient cognitive resources (e.g., higher working memory capacities), further knowledge (e.g., background/prior knowledge, experience, and expertise), special skills (e.g., higher reading skills), or for those that were generally high achieving (e.g., McNamara et al., 1996; Kalyuga et al., 2001; McDaniel et al., 2002; Carpenter et al., 2016; Lehmann et al., 2016). McDaniel et al. (2002) thereby argued that even when learners can correctly solve difficult generation tasks, this consumed a

lot of their processing capacities. This is why only more able readers—and not less able readers—benefitted from generation tasks: Only these learners still had cognitive capacities left to further process and deeper encode the generated information after solving the difficult tasks. Notably, these findings and argumentations indicate that desirable difficulties—and especially tests—have to be difficult, demanding, and taxing to be beneficial but that learners must simultaneously be sufficiently equipped to master these posed challenges, must possess the skills to successfully respond to the difficult tasks and to successfully retrieve information, and must be able to muster the needed increased effort (e.g., Richland et al., 2005; Bjork and Bjork, 2011, 2019; Kornell et al., 2011; Alter et al., 2013; Oppenheimer and Alter, 2014; Rowland, 2014; Karpicke, 2017; Kaiser et al., 2018). This, however, may not prove possible for every learner—but should apply to leaners with higher intelligence.

Tests and Intelligence

Intelligence has often been shown to be one of the strongest predictors for long-term learning, information retrieval, or academic achievement, and it is also argued to be especially valuable and predictive for difficult and stimulating learning environments and complex materials (e.g., Gottfredson, 1997; Kuncel et al., 2004; Fergusson et al., 2005; Bornstein et al., 2013; Roth et al., 2015; Stadler et al., 2015; Stern, 2015, 2017; Strenze, 2015). Moreover, intelligence is even defined as the ability to learn, to reason, and to solve problems and has also often been found to be associated with successful information processing, successful retrieval from long-term memory, and higher working memory capacities (see, e.g., Gottfredson, 1997; Sternberg, 1997; Oberauer et al., 2005; Bornstein et al., 2013; Stern, 2015, 2017; Wang et al., 2017). Hence, taken together, higher intelligence is not only generally important for long-term learning outcomes but also seems to be fundamental for tests to be actually beneficial and for learners to be actually able to reap those benefits. Thus, intelligence should moderate the beneficial effects of tests, insofar as that especially learners with sufficient cognitive abilities and higher intelligence should benefit from desirable difficulties and tests, particularly when learning with complex and curricular materials: Such learners should be able to successfully retrieve, further process, and understand the learned information and to manage such difficult tests without being cognitively overwhelmed—even after working on difficult and cognitive capacities reducing tasks (e.g., Kalyuga et al., 2001; McDaniel et al., 2002; Lehmann et al., 2016). Two previous studies found supporting evidence for the assumption that intelligence moderates the beneficial effects of tests: First, a study from Minear et al. (2018) yielded that higher fluid intelligence increased the positive effects of tests for difficult, as opposed to easy, information (regarding Swahili-English word pairs; learners with lower fluid intelligence showed the reverse effect). Second, Wenzel and Reinhard (2019) found that only at least averagely intelligent learners achieved higher long-term learning in a test condition compared to averagely intelligent learners in a re-reading control condition. Relatively intelligent learners (intelligence one standard deviation above mean) profited even more from difficult tests (Wenzel and Reinhard, 2019). Hence, these argumentations and findings imply that special prerequisites, like average or higher intelligence, must be given so that learners can even reap the benefits of tests. However, contrary findings also exist (showing different or no interactions between intelligence and the effectiveness of tests, e.g., Brewer and Unsworth, 2012; Robey, 2017), so that further work is still valuable.

Interestingly, the findings of Wenzel and Reinhard (2019) also highlighted that relatively unintelligent learners (intelligence one standard deviation below mean)—albeit they indulged in more effort and suffered a more strenuous and demanding way of learning—did not outperform less intelligent learners that instead studied with easier, more fluent, and less demanding re-reading tasks. Thus, the learning outcomes of less intelligent learners in both learning conditions did not differ from each other, whereas learners' subjective experiences and perceptions during learning should have differed strongly. This in turn raises the question if further factors additionally to or beyond long-term learning must be considered when contemplating whether or not to apply tests in school or university settings. For instance, difficult learning tasks were previously shown to increase perceptions of threat or anxiety, experiencing difficulties as well as giving incorrect answers was found to feed negatively into self-perceptions, and performing poorly increased stress perceptions (e.g., O'Neil et al., 1969; Schunk and Gaa, 1981; Sarason and Sarason, 1990). Difficult learning tasks and tasks that require more effort, more time, and more workload were additionally often perceived as more stress-inducing compared to easier tasks (e.g., Kausar, 2010). Thus, tests might result in negative (side) effects like increased stress perceptions (which would be especially undesirable if the respective learners did not even profit from taking such tests).

Tests and Perceptions of Stress or Anxiety

According to the *transactional theory of stress* (e.g., Lazarus and Folkman, 1987), perceptions of stress or anxiety arise when working on tasks (or when being in situations) that are perceived as threatening instead of challenging and in which individuals think that they do not possess enough resources or enough cognitive abilities to cope with the posed demands. Perceived imbalances between difficult tasks and learners' own capabilities or resources also result in stress perceptions (see, e.g., McGrath, 1970; Lazarus, 1990; Kausar, 2010). Unsurprisingly, most students experience test situations, especially (graded) final high-stake tests, (summative) exams, or (competitive) school entrance examinations, as stressful, pressuring, and unpleasant (e.g., Sarason, 1984; Beilock, 2008; Bradley et al., 2010; Jamieson et al., 2016; Leiner et al., 2018). It was also observed that the majority of students' academic stress stems from taking and studying for exams and from getting examination results (see, e.g., Abouserie, 1994). However, apart from such (graded) examinations, even tests solely used as learning situations might be stress-or anxiety-inducing—because tests as desirable difficulties must even per definition be challenging, effortful, and difficult, and might thus be perceived as overwhelming. In line with these assumptions, Hinze and Rapp (2014) conducted a laboratory study using science texts as study materials and

applied re-reading tasks, low-stakes learning tests, or high-stakes learning tests. Stakes were thereby operationalized through instructions given before the learning tests stating that monetary rewards for the learner and a fictive partner were either independent of learners' later final test results or dependent of their later final test results. The authors found that even low-stakes tests led to more immediate feelings of pressure than re-reading tasks and that high-stakes tests further led to more state anxiety than low-stakes tests and re-reading tasks (notably, these results were independent of participants' trait anxiety and there were no interactions between the learning condition and trait variables, Hinze and Rapp, 2014). Another laboratory study also found that learning situations including a short test (on mathematical concepts and materials) were evaluated as more negative and as more stress-and anxiety-inducing than learning situations including a reading control task (these findings were also independent of participants' trait stress or trait anxiety; Wenzel and Reinhard, 2021). Interestingly, contrary results were also found (see, e.g., Agarwal et al., 2014; Nyroos et al., 2016) and even though these can be explained due to methodological differences, replications are still advantageous. Apart from that, it is furthermore possible that these effects of tests on stress perceptions do not arise for learners with higher intelligence and that intelligence might moderate these negative effects.

Intelligence and Perceptions of Stress or Anxiety

Because learners with higher intelligence should generally be able to solve difficult tasks and to answer more test questions successfully, they should, in turn, perceive tests as less threatening, less stressful, less difficult, less overwhelming, and thus as more manageable than learners with lower intelligence. In line with these assumptions, previous work showed that cognitive abilities were negatively correlated to situational stress experiences, math anxiety, state anxiety, and to ratings of difficulty of varying learning tasks (e.g., Efklides et al., 1997; LePine et al., 2004; Abín et al., 2020). Students that were extremely high-achieving in mathematics were also less math anxious, were more motivated, had more self-efficiency, and reported more enjoyment while learning (e.g., García et al., 2016). A study from Goetz et al. (2007) fittingly yielded that emotions experienced by school students during a mathematics achievement test differed based on their abstract reasoning abilities: Anger and anxiety were more prominent for students with lower abilities, whereas enjoyment was more prominent for students with higher abilities. However, if stress nonetheless arises due to tests, such generally unpleasant perceptions are also associated with even further detrimental effects and lower learning outcomes.

Effects of Stress and Anxiety on Learning Outcomes

For instance, higher stress and anxiety were often found to be linked to lower motivation to learn, more errors, lack of concentration, disruptions in attention, higher cognitive load, and reduced effort and persistence while learning (e.g., LePine et al., 2004; Chen and Chang, 2009; Kurebayashi et al., 2012). Anxiety and stress were also negatively correlated with cognitive information processing, the effectiveness of retrieval practice, learning outcomes, academic achievement, and learners (test) performance—especially as the tasks, test questions, or information become more complex, more cognitive demanding, and more difficult (e.g., Hembree, 1988; Seipp, 1991; Struthers et al., 2000; Cassady, 2004a,b; Eysenck et al., 2007; Beilock, 2008; Chen and Chang, 2009; Khan et al., 2013; Sotardi et al., 2020). Hence, stress and anxiety were generally shown to have detrimental effects on learning outcomes but should further also negatively impact the normally beneficial effects of tests. In line with these assumptions, Mok and Chan (2016) found that highly test anxious participants in a learning test condition did not outperform participants in a re-reading control condition. Thus, there were no benefits of tests for highly anxious participants. Similar results were found by Hinze and Rapp (2014): High-stakes learning tests (operationalized through stating that monetary rewards were dependent of participants later final test results) increased pressure and state anxiety directly before the learning tests, which in turn decreased the benefits of these tests regarding later long-term learning. Only participants in a low-stakes learning test condition (in which monetary incentives were not stated to be dependent of participants' test results) outperformed participants in the re-reading control condition. Hence, acute stress perceptions might mediate the beneficial effects of tests, insofar as that higher stress might partly diminish or even completely erase the beneficial effects of tests on long-term learning. Theoretically, such detrimental effects of stress on learning outcomes and on beneficial effects of tests are assumed to arise because stress and anxiety lead to worries and cognitive interference indicated by intrusive, distracting, and irrelevant thoughts. These, in turn, disrupt task-specific information processing, interfere with cognitive processes, impair retrieval, and divert the needed attention and focus away from the learned information, thereby depleting cognitive capacities and storage and processing resources: These consumed resources and capacities would otherwise have been needed for retrieving information, for successfully answering test questions, and for further processing, encoding, or decoding of these information (see, e.g., *attentional control theory, cognitive interference model, distraction theories, processing efficiency theory, and retrieval disruption hypothesis*; Eysenck and Calvo, 1992; Ashcraft and Krause, 2007; Eysenck et al., 2007; Hinze and Rapp, 2014; Sarason, 1984; Tse and Pu, 2012; however, contrary results and contrary theories also exist, showing, for instance, positive linear effects of stress on learning outcomes or non-linear/inverted U-shaped relations of anxiety and performance; see, e.g., LePine et al., 2004; Keeley et al., 2008; Sung et al., 2016). Notably, such detrimental effects of acute stress and anxiety on learning might again be less pronounced for learners with higher compared to learners with lower intelligence. Thus, intelligence might moderate these detrimental effects.

Intelligence and Detrimental Effects of Stress and Anxiety

Because higher intelligence is generally linked to better information processing, higher (working memory) capacities, and better retrieval from long-term memory, learning outcomes of more intelligent learners should not be harmed (as strongly) by stress perceptions, worry, or reduced cognitive capacities compared to learning outcomes of less intelligent learners (e.g., Oberauer et al., 2005; Stern, 2015, 2017; Wang et al., 2017). Thus, such learners should still possess enough resources and capacities to successfully work on difficult tasks and to further process the retrieved and studied information even after perceiving stress. In line with this, researchers assumed that higher domain-specific abilities or extra processing resources should be able to compensate detrimental effects on learners' initial acquisition of information and on their later learning outcomes caused by stress and anxiety (e.g., Tobias, 1984; Naveh-Benjamin, 1991; Eysenck and Calvo, 1992; Eysenck et al., 2007). Fittingly, a study from Tse and Pu (2012) found that less effective and less successful retrieval practice caused by higher test anxiety could be compensated by higher working memory capacities. Thus, anxiety had only detrimental effects for learners with lower working memory capacities (see also Ashcraft and Krause, 2007; Johnson and Gronlund, 2009; Owens et al., 2014; for contrary results, see Beilock, 2008). Previously conducted work also yielded that cognitive abilities had a buffering effect for negative consequences of distraction, insofar as that distraction only had a detrimental effect on (exam) performance for lower ability learners but did not decrease performance of higher ability learners (Reeve et al., 2014). It was furthermore shown that (fluid) intelligence moderated the impact of state anxiety on working memory functioning: The negative impact of state anxiety on working memory functioning was shown to diminish with higher intelligence and anxiety only negatively affected working memory for learners with intelligence below median (Chuderski, 2014).

The Present Research

Taken together, the present research simultaneously focused on tests as desirable difficulties, their beneficial effects on later learning outcomes, and their negative effects on stress perceptions. We further focused on detrimental effects of increased stress on later learning outcomes and on the normally beneficial effects of tests. Moreover, we also explored learners' intelligence as a potential prerequisite for beneficial effects of tests as well as potentially moderating effects of intelligence: Higher intelligence should increase beneficial effects of tests on later learning outcomes, decrease stress perceptions caused by tests, and reduce detrimental effects of stress on learning.

Following the in the Introduction presented empirical and theoretical argumentations, we thereby suppose the following hypotheses (see **Figure 1** for a graphical depiction). For a better comprehensibility, we want to sort the hypotheses according to main and interaction effects: First, we assume that tests, compared to re-reading tasks, result in beneficial effects on later learning outcomes: Thus, a test condition should lead to higher later learning outcomes than a re-reading control condition (*Hypothesis 1*). Nonetheless, working on tests should also increase acute stress perceptions compared to working on the re-reading task (*Hypothesis 2*). In turn, such acute stress perceptions were assumed to be negatively correlated with participants later learning outcomes (*Hypothesis 3*). In that regard, we assumed that acute stress perceptions would mediate the effect of the learning condition (and thus the beneficial effects of tests) on later learning: Higher stress perceptions caused by tests should be linked with reductions of the normally beneficial effects of tests on later learning outcomes. Moreover, we assume intelligence to be positively correlated with later learning outcomes (*Hypothesis 4*).

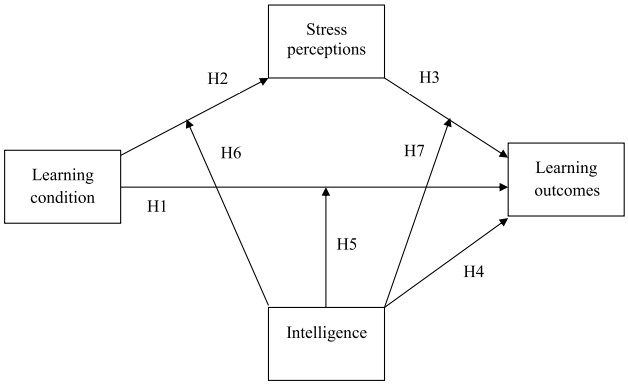


FIGURE 1 | Graphical depiction of the different variables and the assumed hypotheses. The learning condition consists of a re-reading control condition (0) and a test condition (1).

We also assumed the following three interaction effects: First, we assumed that the beneficial effects of tests on later learning outcomes should be moderated by participants intelligence: Beneficial effects should be stronger for more intelligent participants and weaker for less intelligent participants (*Hypothesis 5*). Second, the negative effects of tests on stress perceptions should also be moderated by intelligence: More intelligent participants should perceive less acute stress when learning with a test than less intelligent participants in the test condition (*Hypothesis 6*). Third, the detrimental effects of stress perceptions on later learning outcomes should also be moderated by intelligence: Later learning outcomes of more intelligent participants should be less harmed by stress perceptions than later learning outcomes of less intelligent participants (*Hypothesis 7*).

To test these hypotheses, we conducted a laboratory study consisting of two sessions. We therefore designed a realistic learning situation that could be easily transferred to actual universities or schools. We used, for instance, complex and curricular learning materials that are actually applied in university courses. Thus, we tried to replicate the often found beneficial effect of tests (compared to easier and more passive re-reading control tasks) for difficult and realistic materials. We also conducted a short learning test, including varying test questions formats (e.g., short-answer and multiple-choice questions) that students should often encounter in their university lives (e.g., at

the end of textbook chapters, in examinations, …). Moreover, to reliably investigate whether learning tests actually lead to stress perceptions, we devised an extremely low-stakes learning test situation that still resembled an actual university course as closely as possible. Hence, we did not want to experimentally manipulate stress but wanted to observe if stress perceptions would even occur in virtually pressure-less learning situations that either include a short test task or a re-reading task. Fittingly, we only instructed participants to do their best while learning and did include neither monetary rewards (see, e.g., Hinze and Rapp, 2014) nor grades (see, e.g., Khanna, 2015) as further incentives that might influence their perceptions and evaluations of these learning tests. This also ensured that our laboratory learning situation would resemble a typical learning situation in university or school settings. To further ensure that the test task would be without stakes or artificial stressors, we avoided using learning materials that might be stress-or anxiety inducing in themselves (like mathematical or statistical information; see, e.g., Wenzel and Reinhard, 2021) and applied a test in which participant did not even have to say their answers out loud in front of their peers (contrary to Wenzel and Reinhard, 2021; see also England et al., 2017). To adequately assess participants stress perceptions caused by the learning situation, we measured their state stress directly after they completed the respective learning task and explicitly instructed them to refer to their perceptions and experiences while learning (contrary to previous work where stress was assessed, for instance, before participants worked on the respective tests, after the tests but with a longer delay, or even retrospectively at the end of the academic year; see, e.g., Agarwal et al., 2014; Hinze and Rapp, 2014; Nyroos et al., 2016). Finally, we must note that our work was planned and conducted shortly before the onset of the COVID-19 pandemic. Therefore, our theoretical and methodological considerations mostly focused on conventional learning settings or conventional learning modalities that were rather typical for our respective university before the restrictions due to COVID-19 were implemented. This includes, for instance, face-to-face learning situations in which students learn alongside their peers with a lecturer present as well as directly in-class implemented learning tasks (see, e.g., Yang et al., 2021, for the benefits of supervised in-classroom tests compared to tests administered outside of classrooms). Hence, our laboratory setting was intended to mirror a typical learning situation before most education was transferred to distance e-learning.

MATERIALS AND METHODS

Participants

Power was set to 0.90, and sample size was calculated to detect a medium effect ($f = 0.25$).[1] Using G*Power (Faul et al., 2009), a power analysis revealed a needed sample size of $N = 171$ to detect a significant effect (alpha level of 0.05)—given there is an effect (regrettably, we later realized that—following the argumentation of Blake and Gangestad (2020)—this calculation would have already resulted in an underpowered sample size regarding the assumed interaction effects). Unfortunately, due to the COVID-19-outbreak and later lock-down restrictions, we also had to stop our recruitment and could not continue to collect data in the laboratory (this, in turn, further drastically reduced the power of our work, especially regarding the assumed interaction effects that are extremely underpowered). Due to this stop of our recruitment, our sample consisted of only 91 participants, from which two participants had to be excluded because they did not participate in both sessions of the study. Hence, our final sample consisted of $N = 89$ participants ($M_{age} = 24.18$, $SD_{age} = 6.25$, range: 18–48; 70.8% female; 85.4% German native speakers). Of these, 96.7% were students at a German university. Seventy-three of them (82.00%) studied psychology, and the remaining studied, among others, architecture, education, philosophy, social science, languages, and politics. Each participant was randomly assigned to one of the two between-subjects learning conditions: the re-reading control condition ($n = 47$) or the test condition ($n = 42$). Before starting, each participant had to provide their approval through reading and agreeing to a written informed consent. The study was conducted in full accordance with the Ethical Guidelines of the DGPs and the APA, and the funded project was approved by the Ethics Committee affiliated with the funding source.

Procedure

Up to seven participants could simultaneously take part in our study. On average, 3.83 students participated simultaneously ($SD = 1.97$, range = 1–7). For less diversion and more anonymity, each participant sat in a workplace with dividers in front of a computer. All tasks were complete on this computer. In general, participants arrived together, started the study together, and worked simultaneously on the specific tasks but did not directly communicate with each other while undergoing the study and while learning. Apart from a brief welcome from the experimenter, short instructions when different tasks were supposed to start and stop, and a short farewell (all oral instructions were read out loud from standardized texts), all materials and all instructions were presented on the computer. The experimenter (the first author) otherwise only stopped the time for time-limited tasks, made sure that these time limits were met, and monitored that participants generally adhered to the instructions (e.g., the experimenter sometimes reminded participants to further work on the specific learning tasks if participants had stopped working although they still had time left for studying).

Session 1

After a brief welcome and after reading and agreeing to the written informed consent, participants' demographic measures were assessed (e.g., age, gender, occupational status, native language, ethnicity, field of study, and graduation grade). Thereafter, we measured an intelligence estimate using a 3-min

[1]Our study was pre-registered by AsPredicted (see https://aspredicted.org/dm7rd.pdf).

intelligence screening (*mini-q*; Baudson and Preckel, 2015; based on Baddeleys *verbal reasoning*, Baddeley, 1968; further: *intelligence-estimate*). The mini-q is a reliable and valid screening instrument for general (fluid) cognitive abilities that accurately assesses speeded reasoning as a conglomerate of reasoning, abstract thinking, and processing speed (Baudson and Preckel, 2015). The mini-q includes 64 tasks that each consist of a statement describing three geometrical figures (square, triangle, and circle) that participants have to declare as right or wrong (for two example items, see Baudson and Preckel, 2015) and have 3 min to solve as many of the tasks as possible.[2] Using a standard table including a representative adult sample, the sums of correctly solved tasks can then be transformed to estimations of intelligence scores ($M = 100$, $SD = 15$). Participants were generally instructed to try their best while working on these task. To ensure that our instructions would not frame the task as needles pressuring or stressful, we correctly described that the task focused on participants reasoning and abstract thinking abilities but did not explicitly highlight that it thereby also serves as an intelligence-estimate. This was done because previous work sometimes induced stress perceptions by explicitly presenting tasks as intelligence tests or by using instructions that generally increase participants' expectations of having to work on demanding or threatening intelligence tests (see, e.g., Kimmel and Bevill, 1985; Zeidner, 1998).

Before the learning phase started, we then informed participants that we wanted to explore the effectiveness of different learning tasks, which is why it would be important that they give their best while learning and that they should imagine to be studying for one of their actual university courses. We also reminded them that the ability to quickly and successfully learn new information is extremely advantageous in their everyday university lives and asked them to learn as intensively as they normally would. Participants were also informed that they would, 1 week later, be charged with taking a final test covering the learned information. The learning materials consisted of one textbook chapter describing the brain's lateralization based on a standard introductory textbook that is often adopted for university courses in biopsychology (Pinel and Pauli, 2012). Thus, the learning material was difficult, complex, and curricular. Before participants initially read the text, we assessed their prior knowledge regarding this topic to check if it differed between participants in the two learning conditions. We thereby implemented three open-ended questions (e.g., *Which function is linked to the Broca area?*) that participants answered within 3 min.

In the following *first learning phase*, all participants once read the three textbook pages concerning the brain's lateralization as an initial study opportunity. They were therefore given about 10 min. For the subsequent 10 min of the *second learning phase*, each participant was then (*via* the computer they worked on) randomly and individually assigned to either the re-reading control condition or to the test condition.

Re-Reading Control Condition

In the re-reading control condition, participants were again presented with the textbook chapter. They were instructed to read the text as often as they wanted in the given time and to learn, understand, and memorize the information.

Test Condition

In the test condition, participants were presented with a learning test inquiring different aspects of the previously read textbook. The test consisted of 17 questions. These were multiple-choice questions and open-ended questions, which required both short answers consisting of single words or bullet points as well as longer, more detailed answers (participants could gain up to 2 points per correct answer; a maximum of 20 points could be gained; for examples, see **Appendix A**).

Following, participants state stress caused by the learning condition was measured with the German version of the Perceived Stress Questionnaire (PSQ; Fliege et al., 2001; based on Levenstein et al., 1993) using 20 items ($\alpha = 0.89$; e.g., *You felt tense*) on a four-point Likert-like scale from one (*almost never*) to four (*usually*). To assess participants immediate stress perceptions, they were explicitly instructed to refer their ratings to their perceptions and experiences during the just finished second learning phase.

Participants then answered some manipulation check questions regarding the second learning phase, e.g., regarding the difficultly, strenuousness, or helpfulness of the learning task, their assumed success, as well as their evaluations of the second learning phase as negative/positive and challenging/threatening (e.g., *How difficult did you find working on the second learning phase?* one (*very easy*) to five (*very difficult*); see **Appendix A** for all 6 manipulation check questions). Thereafter, participants in the test condition received feedback in form of an answer sheet displaying the correct answers to the test questions. Finally, participants were asked if they had already known the learning materials or the applied intelligence screening and were instructed not to study the learned materials in the meantime.

Session 2

In the second session (1 week after Session 1; $M_{days} = 7.12$, $SD_{days} = 0.50$, range: 7–10), participants later learning outcomes were assessed. Therefore, participants were required to work on a final test for 10 min. The final test included 21 questions (participants could gain up to 2 points per correct answer; a maximum of 27 points could be gained). In line with the learning test in Session 1, the final test consisted of multiple-choice and open-ended questions. Eight of the final test questions were identical to questions previously used in the learning test, while seven of them were slightly changed to assess transfer. The remaining six final test questions asked about information that were part of the read textbook chapter but had not been previously implemented in the learning test in Session 1.

[2]The procedure of the mini-q was—in accordance with Prof. Dr. Tanja Baudson—slightly adapted: Instead of letting participants solve all tasks without a time limit and to then use the number of correctly solved tasks at exactly 3 min, we directly terminated the measurement after 3 min.

In the end, participants were asked if they had re-studied the learning materials in the interim. They were then shortly debriefed and received the opportunity to take part in a raffle for a total of 200 Euro. Psychology students could alternatively earn course credit.

RESULTS

Participants' age, gender distribution, native language distribution, graduation note, the number of students that participated simultaneously, the time lag between Sessions 1 and 2, participants' intelligence-estimate scores, and their prior knowledge did not significantly differ between the test condition and the re-reading control condition (all $ps \geq 0.163$). This indicated that the random distribution of participants to the two conditions had been successful. Comparing the manipulation check questions between participants in the test condition and participants in the re-reading control condition indicated that the manipulation of the conditions had also been successful: Most important, participants in the test condition rated the learning situation as significantly more difficult than participants in the re-reading control condition, $M_{\text{re-reading}} = 2.11$, $SD_{\text{re-reading}} = 0.96$, $M_{\text{test}} = 2.90$, $SD_{\text{test}} = 1.12$, $t(87) = -3.62$, $p = 0.001$, $d = -0.76$ (95% CI[−1.20; −0.32]). The effect size can be classified as medium to high. The test condition was also evaluated as slightly more challenging than the re-reading control condition, $M_{\text{re-reading}} = 2.74$, $SD_{\text{re-reading}} = 0.57$, $M_{\text{test}} = 2.26$, $SD_{\text{test}} = 0.83$, $t(87) = 3.23$, $p = 0.002$, $d = 0.68$ (95% CI[0.25; 1.11]). There were no significant differences between ratings of strenuousness, helpfulness, overall (positive or negative) evaluation, and successfulness of the two learning conditions (all $ps \geq 0.081$).

Descriptively, participants achieved on average an intelligence-estimate score of 112.03 ($SD = 16.21$, range: 73–154). Their average state stress score was 2.09 ($SD = 0.52$, range: 1.20–3.70). Considering the final test measuring their later learning outcomes, participants were on average able to give 13.84 of 27 (51.26%) correct answers ($SD = 4.33$, range: 4–24).

To test Hypothesis 1, we conducted a *t*-test to compare participants later learning outcomes in both learning conditions: $M_{\text{re-reading}} = 12.87$, $SD_{\text{re-reading}} = 4.17$, $M_{\text{test}} = 14.93$, $SD_{\text{test}} = 4.30$, $t(87) = -2.29$, $p = 0.025$, $d = -0.49$ (95% CI[−0.92; −0.06]). As assumed, participants in the test condition answered more final test questions correctly than participants in the re-reading control condition, serving as first support for Hypothesis 1. The size of this effect can be interpreted as medium.

Following, we conducted another *t*-test to compare participants' acute stress perceptions in both learning conditions to test Hypothesis 2: $M_{\text{re-reading}} = 1.99$, $SD_{\text{re-reading}} = 0.49$, $M_{\text{test}} = 2.21$, $SD_{\text{test}} = 0.52$, $t(87) = -2.04$, $p = 0.045$, $d = -0.44$ (95% CI[−0.87; −0.01]). Supporting Hypothesis 2, participants in the test condition perceived more state stress during and immediately after the learning situation compared to participants in the re-reading control condition. The size of this effect can be classified as small to medium.

In turn, such stress perceptions were significantly and negatively correlated with later learning outcomes {$r = -0.26$ (95% CI[−0.44; −0.06]), $p = 0.014$},[3] showing a small to medium correlation. Thus, higher stress perceptions were linked to lower later learning outcomes indicated by fewer correctly solved final test questions. This served as first support for Hypothesis 3.

To test whether the beneficial effects of tests on later learning outcomes were mediated by participants acute stress perceptions, we then ran a mediation analysis[4] with Process (model 4; Hayes, 2018). Thus, we tested direct effects of the learning condition on participants later learning outcomes and indirect effects of the learning condition on participants later learning outcomes *via* state stress (all predictors and the potential mediator were z-standardized; see **Figure 1** for a graphical illustration of these assumed relations and our hypotheses). The learning condition significantly predicted participants perceived stress during the learning situation (path a), $B = 0.43$, $SE = 0.21$, $t(87) = 2.03$, $p = 0.045$. Thus, tests increased acute stress perceptions, which served as further evidence for Hypothesis 2. In turn, such state stress predicted participants later learning outcomes (path b), $B = -1.41$, $SE = 0.39$, $t(86) = -3.60$, $p = 0.001$. Thus, higher stress perceptions were linked to lower later learning outcomes, serving as further evidence for Hypothesis 3. We also found a significant total effect (path c) of the learning condition on later learning outcomes, $B = 2.06$, $SE = 0.90$, $t(87) = 2.28$, $p = 0.025$. The direct effect (path c') of the learning condition on later learning outcomes (when simultaneously controlling for participants' stress perceptions) was also significant, $B = 2.66$, $SE = 0.88$, $t(86) = 3.04$, $p = 0.003$. Thus, we found the assumed beneficial effects of tests on later learning, which served as further evidence for Hypothesis 1. Moreover, the indirect effect of the learning condition on participants later learning outcomes *via* state stress was also significant (path a x path b), $B = -0.60$, 95% CI[−1.47; −0.04]. Notably, the direct effect was stronger than the total effect, showing that controlling for participants' state stress increased the beneficial effects of the test condition. This indicated that state stress is not a mediator but a suppressor of the effect of the learning condition on later learning outcomes.

Furthermore, correlational analyses then showed that participants later learning outcomes were significantly correlated

[3]Exploratively conducted (hierarchical regression) analyses further supported—at least concerning this study and this sample—the assumed linear (instead of a polynominal/non-linear) relation between participants stress perceptions and their later learning outcomes: Neither a regression model assuming a quadratic nor a regression model assuming a cubic link between stress and later learning outcomes was able to explain more variance than a model assuming a linear relation [both $ps \geq 0.342$; see also Sotardi et al., 2020 regarding this approach and similar findings].

[4]In line with typically used wordings regarding mediation analyses, we will also refer to the regression analysis testing the effect of the predictor (learning condition) on the potential mediator (stress perceptions) as *path a* and to the regression analysis testing the effect of the mediator (stress perceptions) on the criterion (learning outcomes) as *path b*. We also refer to the effect of the predictor (learning condition) without controlling for the mediator (stress perceptions) on the criterion (learning outcomes) as *path c* (*total effect*) and to the effect of the predictor (learning condition) on the criterion (learning outcomes) while controlling for the potential mediator (stress perceptions) as *path c'* (*direct effect*). We also refer to the indirect effect of the predictor (learning condition) on the criterion (learning outcomes) *via* the mediator (stress perceptions) as *path a x path b*.

with their intelligence-estimates {$r = 0.34$ (95% CI[0.14;0.51]), $p = 0.001$, showing a medium correlation}. This served as first support for Hypothesis 4. Interestingly, the intelligence-estimate was also significantly—and negatively—correlated with participants state stress {$r = -0.39$ (95% CI[−0.55; −0.20]), $p < 0.001$, showing a medium correlation}.

Finally, we conducted a moderated mediation analysis (Process, model 59; Hayes, 2018) to test all hypotheses—including the three assumed interaction effects (Hypotheses 5, 6, and 7)—simultaneously in a single statistical model (all predictors, the mediator, and the moderator were z-standardized; see **Figure 1** for a graphical illustration of these assumed relations and our hypotheses). Because not all requirements were fulfilled (homoscedasticity was not given for one path of the mediation analysis, Breusch-Pagan test: $p = 0.031$), we ran this analysis with heteroscedasticity robust standard errors imbedded in Process. Again, the learning condition significantly predicted participants perceived stress during the learning situation (path a), $B = 0.40$, $SE = 0.20$, $t(85) = 2.05$, $p = 0.043$. The intelligence-estimate was also a significant predictor for such stress perceptions, $B = -0.34$, $SE = 0.15$, $t(85) = -2.26$, $p = 0.027$. However, the intelligence-estimate did not moderate this negative effect of the learning condition on stress perceptions (learning condition*intelligence-estimate), $B = -0.10$, $SE = 0.19$, $t(85) = -0.55$, $p = 0.586$. Taken together, tests led to more acute stress perceptions than the re-reading control task, which again supported Hypothesis 2. Notably, although higher intelligence was generally linked to lower stress perceptions, the effect of the learning condition on stress perceptions was not moderated by the intelligence-estimate, thereby not supporting Hypothesis 5. Moreover, state stress, in turn, again predicted participants later learning outcomes (path b), $B = -1.01$, $SE = 0.50$, $t(83) = -2.04$, $p = 0.045$. The intelligence-estimate was, contrary to the previously conducted correlational analysis, not a significant predictor for later learning outcomes, $B = 1.16$, $SE = 0.67$, $t(83) = 1.73$, $p = 0.088$. The intelligence-estimate did also not moderate the detrimental effect of stress perceptions on later learning outcomes (stress perceptions*intelligence-estimate), $B = -0.12$, $SE = 0.54$, $t(83) = -0.22$, $p = 0.829$. Thus, higher stress perceptions were again linked to lower later learning outcomes, which again supported Hypothesis 3. However, intelligence neither predicted later learning outcomes nor moderated the detrimental effect of stress on later learning outcomes, hence, neither supporting Hypothesis 4 nor Hypothesis 6. Furthermore, there was a significant direct effect (path c') of the learning condition on later learning outcomes, $B = 2.54$, $SE = 0.85$, $t(83) = 2.98$, $p = 0.004$. This effect was also not moderated by the intelligence-estimate (learning condition*intelligence-estimate), $B = -0.10$, $SE = 1.01$, $t(83) = -0.10$ $p = 0.919$. These findings again showed that tests were more beneficial for participants later learning outcomes than the re-reading control task and that this beneficial effect was independent of participants intelligence. This again supported Hypothesis 1 but not Hypothesis 7. The indirect effect of the learning condition on later learning outcomes *via* stress perceptions did also not differ depending on participants' intelligence-estimates.

Exploratory Analyses

Exploratory analyses can be found in **Appendix B**. These include, for instance, analyses focusing separately on the three different types of final test questions indicating later learning outcomes described in the methods section. We also depict correlations among participant ratings of the manipulation check questions (assessing their perceptions and evaluations of the two learning conditions) and participants stress perceptions.[5]

DISCUSSION

The present work was conducted to simultaneously test linkages among (learning) tests, acute stress perceptions, intelligence, and later learning outcomes (see **Figure 1** for a graphical overview of our hypotheses). Addressing these linkages and testing our hypotheses is extremely relevant before tests—as potentially double-edged swords—are used in university and school settings. Summarizing, our results supported all assumed main effects (most effect sizes can thereby be categorized as small to medium) but none of the assumed interaction effects. In more detail, our data yielded that tests led to higher later learning outcomes 1 week after the learning phase compared to the re-reading control condition. This fits the literature mentioned in the Introduction and again shows the benefits of applying tests as difficult learning tasks (e.g., Rowland, 2014; Adesope et al., 2017; Yang et al., 2021). However, also in line with our assumptions and the in the Introduction cited literature (e.g., Hinze and Rapp, 2014; Wenzel and Reinhard, 2021), the test condition also increased participants acute stress perceptions during and directly after learning compared to the re-reading condition. Although the descriptive statistics of stress perceptions were not extremely high (midpoint of the scale = 2.00, $M_{\text{re-reading}} = 1.99$, $M_{\text{test}} = 2.21$) and the size of the effect was only small to medium, our results showed that even low-stakes learning tests were perceived as more demanding, more threatening, and more stressful than re-reading of previously studied materials. In turn, such stress perceptions were then negatively linked to later learning outcomes, thus supporting previous work that also reported detrimental effects of stress and anxiety on learning (e.g., Seipp, 1991; Hinze and Rapp, 2014; Sotardi et al., 2020). Interestingly, such increased stress perceptions served as a suppressor of the beneficial effects of tests on later learning outcomes (a mediation analysis found an indirect effect of the learning condition on long-term learning *via* stress perceptions): The direct effect of the learning condition controlling for stress perceptions was stronger than the total effect of the learning condition without controlling for differences in stress perceptions. Thus, the beneficial low-stakes test increased participants immediate stress perceptions and these triggered stress perceptions were in turn related to decreases of benefits

[5]We would like to thank an anonymous reviewer for this suggestion.

of the test. Hence, although the test condition was still—albeit less—beneficial for later learning outcomes, it was even more effective when individual differences in stress perceptions were controlled for. Furthermore, as has often been shown before (see, e.g., Kuncel et al., 2004; Fergusson et al., 2005), higher intelligence was linked to higher achievement and higher later learning outcomes.[6] Notably, higher intelligence-estimate scores were additionally related to lower stress perceptions in the learning situation. Thus, higher intelligence buffered feelings and perceptions of threat, demands, or pressure—which is also in line with literature cited in the Introduction (see, e.g., Efklides et al., 1997; LePine et al., 2004; Goetz et al., 2007). Nonetheless, intelligence did not moderate any of the main effects found in our study: The three hypotheses concerning interaction effects (learning condition*intelligence-estimate on stress perceptions, learning condition*intelligence-estimate on later learning outcomes, and stress perceptions*intelligence-estimate on later learning outcomes) were not supported by our data.

Two aspects of our sample were probably the main reasons that we were not able to support these hypothesized interaction effects: the intelligence-estimate scores of our participants and the size of our sample. Although the intelligence-estimate scores of our sample were normally distributed, participants had an average intelligence of 112.03 ($SD = 16.21$, range = 73–154), indicating that even the less intelligent participants in our sample were rather intelligent. In comparison, the relatively unintelligent learners that did not benefit from learning tests in the work of Wenzel and Reinhard (2019; Study 2) had intelligence scores lower than 86.39. In our sample, however, only three participants had intelligence scores that were lower than 86 (73, 84, and 85). Thus, we might have not been able to observe interaction effects due to these already relatively high intelligence scores. Even more important was, however, the small sample size of our work: As mentioned in our methods section, the sample size was—due to the COVID-19-outbreak and the resulting stop of our laboratory study—smaller than *a-priori* calculated (and the *a-priori* conducted and pre-registered sample size might erroneously have already been too small regarding potential interaction effects; see, e.g., Blake and Gangestad, 2020). Thus, it is most likely that the interaction effects were not detected because power was not sufficient.

All in all, even though not all our hypotheses were supported and although the sizes of the found effects can mostly be described as medium, our work raised important research issues and aims to serve as a first step to give (empirically well-grounded) advice to lecturers and teachers regarding the application of tests, their prerequisites, and their (positive as well as negative) consequences. Notably, the simultaneous testing of beneficial learning effects of tests, increased stress perceptions as negative (side) effects caused by tests, detrimental effects of such increased stress perceptions, and also potential moderating effects of learners intelligence has, to our knowledge, not been done before. Hence, our study highlights important research issues, uniquely contributes to the research field, and presents findings that are extremely stimulating for future work. Positively, we therefore conducted a laboratory setting that was similar to realistic learning situations in university settings (at least in this respective university and before the outbreak of the COVID-19 pandemic), insofar as that multiple students simultaneously worked on learning tasks with an experimenter present. Participants were thereby only instructed to learn as they typically would and to do their best without giving them further incentives to do well (like, e.g., monetary incentives that are normally not present in university settings). Moreover, the laboratory was set in a university building that hosts offices of lecturers as well as seminar rooms and many participants participated before or after their normal courses—hence, the setting of the study should have strongly resembled a typical university setting. Most important, the applied learning materials were complex and realistic materials that are actually applied in university courses and that are even—at least for most of the psychology students included in our sample—part of their curriculum. Regarding the test condition, we designed a short, realistic, low-stakes test, which included varying test question types (e.g., multiple-choice questions and short-answer questions requiring both shorter and longer answers) as well as varying levels of questions depths (e.g., asking for facts or asking for understanding, transfer, and application of the initially studied information). These test questions should closely resemble questions that are typically posed in university courses or that are included at the end of chapters found in many textbooks. Thus, our findings—indicating a benefit of short learning test that only require 10 min of students' time and that include varying complex test questions and difficult and curricular information—should be applicable and transferable to learning situations in actual universities and should not only be valid in laboratory settings. Hence, in line with previous work, we would advise lecturers to use the last 10 min at the end of their courses to apply test questions concerning the contents of the respective lectures to help increase their students learning outcomes (this could be done, for instance, at the end of all or only some lecturers; see, e.g., Pashler et al., 2007; McDaniel et al., 2011; Iwamoto et al., 2017; Greving and Richter, 2018). Our work also indicates that such tests are beneficial for all university students independent of their intelligence and might, thus, be applied in different courses, different study paths, and for different educational backgrounds. However, our work also highlights negative (side) effects and detrimental effects caused by tests that lectures should consider and keep in mind when designing and using tests. Even though these effects were expected, they are still startling insofar as that the applied test was short, did not focus on excessively stress-inducing materials, and had no consequence for participants' everyday lives. In line with this, participants worked on their own, did not have to say their answers out loud in front of their peers,

[6]Interestingly, intelligence was also positively correlated with participants number of correctly answered test questions in the test condition ($N = 42$, $r = 0.40$, $p = 0.008$). The number of correctly answered test questions was then, in turn, positively correlated with participants long-term learning ($N = 42$, $r = 0.86$, $p < 0.001$) and negatively with their acute stress perceptions ($N = 42$, $r = -0.54$, $p < 0.001$). These findings highlight the importance of students' successfulness while working on difficult learning tests and the importance of their (cognitive) abilities to solve such difficult tasks (see also Richland et al., 2005).

and knew that their results would remain anonymous and that they only had to try their best without fearing consequences due to their performances (on, for instance, monetary incentives, grades, or general evaluations). Thus, although we conducted the test as a low-stakes learning situation in a laboratory setting without manipulating stress perceptions (and without choosing especially stressful tasks or information), the test nonetheless increased stress perceptions. This indicated that these found negative (side) effects of tests might be even more pronounced in actually relevant learning situations in schools or universities. Due to this assumption and due to the observed further detrimental effects of by tests caused stress perceptions on the beneficial effect of test, tests should be conducted as low-stakes and as stressless as possible—to optimize the benefits of tests on learning outcomes as well as to improve learners' experiences and perceptions while learning. Thus, lecturers should try to implement tests that are at most similarly stress-inducing as the tests we applied in this work or try to design tests that are even less pressuring or threatening (without simultaneously reducing the difficulty of the test that is needed for the beneficial long-term learning effects of tests). For instance, previous work indicated that lectures might try to use more gamified learning strategies: Iwamoto et al. (2017), for instance, showed that short tests applied with Kahoot were beneficial for students learning outcomes and were even perceived and rated as positive by the respective students (see also Wang and Tahir, 2020 regarding the application of Kahoot, as well as Mavridis and Tsiatsos, 2017 for the application of game-based tests). The present work furthermore again showed the relevance of (higher) intelligence—albeit, it did not moderate any of the found effects—for cognitive variables like learning outcomes but also for affective variables like emotional reactions to potentially threatening situations. Although learners perceived tests as more stressful independent of their intelligence and although they similarly suffered under decreased learning outcomes due to higher stress perceptions independent of their intelligence, participants with higher intelligence still had some advantages compared to participants with lower intelligence, insofar as that higher intelligence was linked to less stress perceptions in both learning conditions.

Nonetheless, we have to note that our work is not without limitations, which is why the just described indications and applications should be considered with caution until further replications support our findings (especially regarding the conducted analyses testing the assumed interaction effects). Hence, we want to briefly discuss the limitations of our study as well as outlooks and ideas for future work. The most important limitation is, of course, that our sample size was smaller than *a-priori* calculated and that our work was therefore (especially regarding the assumed interaction effects) underpowered. Thus, future studies should in any case replicate our findings with a much bigger sample (see, e.g., Blake and Gangestad, 2020). Additionally, a large proportion (82.00%) of our participants studied psychology and were rather intelligent ($M = 112.03$, $SD = 16.21$). Thus, collecting a generally more diverse sample and a sample with more variance regarding participants' intelligence scores is important for future work and for future replications—to ensure that the resulting findings are generalizable to different samples and to be able to give empirically well-grounded advice to lecturers and learners. The same applies to future replications using different (e.g., longer or multiple) tests, varying learning materials (e.g., regarding information that are definitely part of students curriculum and that are part of later graded examinations), or different (e.g., real university or school) settings. Future work could also focus more closely on potential impacts of different types of test questions on students' perceptions or learning outcomes (see, for instance, **Appendix B** for exploratory analyses separating the in the present work applied three types of test questions). We also think that it would be valuable to conduct replications that try to control more strongly how participants in the re-reading control condition studied—hence, it is important to know if (and how often or how engaged) participants actually re-read the materials or if they simply skimmed through the text. Although the experimenter of our work reminded participants to keep reading if they had obviously stopped reading before the time limit was up, we unfortunately had no way of knowing if participants actually read the text, how often or how intensively they read the text, and if they thereby actually tried to understand and memorize the presented information. Thus, if participants only browsed through the text and did not genuinely re-read the text, this might have further increased the difference between the two learning conditions. Therefore, it would be advantageous if future work could focus even more on the re-reading control condition or if they could apply different, even stronger control conditions (e.g., note-taking). Additionally, longer delays between the learning phase and the later learning assessment would also be valuable to generalize our results found after 1 week to longer delays and to more durable long-term learning effects. Furthermore, the future work could also use different or additional intelligence tests to focus even more on this important individual difference. Although the applied screening instrument serves as a reliable and valid estimation of general cognitive abilities as a conglomerate of reasoning, abstract thinking, and processing speed, it would still be advantageous to test whether the same results would arise when using longer, more general, or more complex intelligence measurements without short time limits. Chuderski (2014), for instance, stated that shorter and timed intelligence tests—which applies to the used intelligence screening—are often very similar to tests assessing working memory capacities. Thus, further replications would be valuable. Fittingly, future studies could also focus more closely on the assumed effects of intelligence on the benefits of tests to further investigate why or how these might arise: Should more intelligent learners, for instance, benefit more from tests because they are able to answers more questions successfully or because they can (independent of their actual success) better and deeper process the retrieved information and the solved answers? Apart from that, future work should also focus on ways to reduce stress perceptions caused by tests to maintain their benefits: For instance, researchers and lecturers could also test the application of emotion regulation techniques, coping strategies, online test formats, or repeated tests, and they could further

prime the beneficial effects of tests or could generally try to modify learners' perceptions of increased effort as helpful and of stressful situations as challenging instead of threatening (e.g., Struthers et al., 2000; Leeming, 2002; Cassady and Gridley, 2005; DeVaney, 2010; McDaniel et al., 2011; Jamieson et al., 2016; Khng, 2017; see also **Table 1** in **Appendix B** for potential starting points regarding linkages among participants evaluations of learning situations and their stress perceptions). Future work could also explore how long-lasting and robust the negative effects of tests on stress perceptions are.

Finally, we would also like to point out that—because our study was conducted slightly before the COVID-19-outbreak and the resulting restriction and thereby triggered changes concerning students daily lives and their learning experiences—findings of replications and future studies might differ due to these interim events: For instance, recent work showed that students had to adjust to remote learning in response to the pandemic and that as a result their achievement goals, engagement, and perceptions of academic success decreased during his time (e.g., Daniels et al., 2021). Orlov et al. (2021) similarly described that students performed, on average, worse during the pandemic than during previous semesters. Concerning students stress perceptions, the results are not that clear: Whereas some studies found that stress and anxiety perceptions generally increased (see, e.g., Limcaoco et al., 2020; Wu et al., 2021; Yang et al., 2021), some work showed that academic stress first increased but then decreased to pre-COVID levels (see, e.g., Charles et al., 2021). Other studies even yielded that studying during COVID-19 had no effects on students' stress perceptions triggered by learning processes (see, e.g., de la Fuente et al., 2021). Zhang and Liu (2021) further showed that students attitudes toward digital learning influenced the levels of distress they experienced due to the COVID-19 pandemic. Hence, although the findings are not consistent, they highlight that it would be valuable to explore if students stress perceptions or experiences and evaluations of tests (especially regarding remote or digital learning tests) changed in the interim and if these changes might impact their effectiveness. Thus, focusing more strongly on e-learning—as the momentarily most prominent form of learning—seems to be extremely relevant. In line with this, the COVID-19 pandemic and the resulting transfer to remote e-learning also illustrated, among others, the importance and general need for more computerized learning strategies, for more technological applications or digital technologies while learning, or for more innovative, interactive, and gamified teaching strategies to successfully adapt to the current situation and to successfully move to online teaching (see, e.g., Adedoyin and Soykan, 2020; Fergus, 2020; König et al., 2020; Sarju, 2020; Muthuprasad et al., 2021; Nieto-Escamez and Roldán-Tapia, 2021; Obrero-Gaitán et al., 2021; Pozo et al., 2021; Yu et al., 2021). Future work could accordingly investigate the effects of new technologies and of digital learning on education in general but specifically on the application of normally beneficial tests. Hence, future work might focus on, among others, computerized learning and testing, automated scoring systems for tests, automated test question generation, the usage of artificial intelligence in learning, AI-based learning assistants, intelligent tutoring systems, or cyber physical systems in general (see, e.g., Park and Choi, 2008; Grimaldi and Karpicke, 2014; Bachir et al., 2019; Matayoshi et al., 2020; Pugh et al., 2020; Schmohl et al., 2020; Nouhan et al., 2021; see also Radanliev et al., 2020, for a literature review of challenges in the application of artificial intelligence in cyber physical systems). It is thus even more important to conduct further work and to obtain more recent data concerning the in this paper identified issues.

CONCLUSION

All in all, our work showed that the application of tests as a desirable difficulty improves later learning outcomes compared to re-reading of the same materials. This applies to curricular and complex learning materials as well as to realistic and difficult test questions and was even independent of participants' intelligence-estimate. However, the application of such beneficial tests also resulted in higher immediate stress perceptions—even though the test was conducted as a short, low-stakes learning situation. This indicates that actual learning situations including tests might lead to even higher stress perceptions. These stress perceptions were, in turn, linked to diminished benefits of tests. More specifically, controlling for such stress perceptions showed that (at least in this sample) the applied test was even more beneficial when it was not perceived as stressful—or at least as only averagely stressful. Moreover, although there were no moderating effects, higher intelligence was again linked to higher learning outcomes and was even associated with lower immediate stress perceptions during the learning situation.

Hence, our work highlighted important research issues and resulted in interesting findings. Nonetheless, future work is still needed to replicate our study with a much bigger and more diverse sample to explore the robustness of the found effects, to generalize our findings, and to be able to give empirically well-grounded recommendations to lecturers. Moreover, future research should take a closer look at potentially moderating effects of intelligence to ascertain if these effects exist or not. Future work could also try to reduce stress perceptions caused by tests.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation

and institutional requirements. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

KW and M-AR contributed to the study conception and design. Material preparation, data collection, and analyses were performed by KW. Funding acquisition and supervision was performed by M-AR. The first draft of the manuscript was written by KW and M-AR, and KW and M-AR commented on previous versions of the manuscript. All authors read and approved the final manuscript.

FUNDING

This research was supported by a LOEWE grant from the Hessian Ministry for Science and the Arts entitled "desirable difficulties; intrinsic cognitive motivation and performance expectancies" awarded to the co-author.

ACKNOWLEDGMENTS

We want to thank Tanja Baudson for sending us the mini-q and for answering our questions regarding its application. We also want to thank Agnes Thurmann for her help with recruiting and data collection.

REFERENCES

Abín, A., Núñez, J. C., Rodríguez, C., Cueli, M., García, T., and Rosário, P. (2020). Predicting mathematics achievement in secondary education: the role of cognitive, motivational, and emotional variables. *Front. Psychol.* 11:876. doi: 10.3389/fpsyg.2020.00876

Abouserie, R. (1994). Sources and levels of stress in relation to locus of control and self esteem in university students. *Educ. Psychol.* 14, 323–330. doi: 10.1080/0144341940140306

Adedoyin, O. B., and Soykan, E. (2020). Covid-19 pandemic and online learning: the challenges and opportunities. *Interact. Learn. Environ.*, 1–13. doi: 10.1080/10494820.2020.1813180, [Epub ahead of print].

Adesope, O. O., Trevisan, D. A., and Sundararajan, N. (2017). Rethinking the use of tests: a meta-analysis of practice testing. *Rev. Educ. Res.* 87, 659–701. doi: 10.3102/0034654316689306

Agarwal, P. K., D'Antonio, L., Roediger, H. L., McDermott, K. B., and McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *J. Appl. Res. Mem. Cogn.* 3, 131–139. doi: 10.1016/j.jarmac.2014.07.002

Alter, A. L., Oppenheimer, D. M., and Epley, N. (2013). Disfluency prompts analytic thinking—But not always greater accuracy: response to. *Cognition* 128, 252–255. doi: 10.1016/j.cognition.2013.01.006

Alter, A. L., Oppenheimer, D. M., Epley, N., and Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. *J. Exp. Psychol. Gen.* 136, 569–576. doi: 10.1037/0096-3445.136.4.569

Ashcraft, M. H., and Krause, J. A. (2007). Working memory, math performance, and math anxiety. *Psychon. Bull. Rev.* 14, 243–248. doi: 10.3758/BF03194059

Bachir, S., Gallon, L., Abenia, A., Aniorté, P., and Exposito, E. (2019, August). "Towards autonomic educational cyber physical systems," in *2019 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*; August 19–13, 2019; (IEEE), 198–1204.

Baddeley, A. D. (1968). A three-minute reasoning test based on grammatical transformation. *Psychon. Sci.* 10, 341–342. doi: 10.3758/BF03331551

Batsell, W. R., Perry, J. L., Hanley, E., and Hostetter, A. B. (2017). Ecological validity of the testing effect: the use of daily quizzes in introductory psychology. *Teach. Psychol.* 44, 18–23. doi: 10.1177/0098628316677492

Baudson, T. G., and Preckel, F. (2015). mini-q: Intelligenzscreening in drei Minuten [mini-q: intelligence screening in three minutes]. *Diagnostica* 62, 182–197. doi: 10.1026/0012-1924/a000150

Beilock, S. L. (2008). Math performance in stressful situations. *Curr. Dir. Psychol. Sci.* 17, 339–343. doi: 10.1111/j.1467-8721.2008.00602.x

Bertsch, S., Pesta, B. J., Wiscott, R., and McDaniel, M. A. (2007). The generation effect: a meta-analytic review. *Mem. Cogn.* 35, 201–210. doi: 10.3758/BF03193441

Bjork, R. A. (1994). "Memory and metamemory considerations in the training of human beings," in *Metacognition: Knowing About Knowing*. eds. J. Metcalfe and A. Shimamura (Cambridge, MA: MIT Press), 185–205.

Bjork, R. A., and Bjork, E. L. (1992). "A new theory of disuse and an old theory of stimulus fluctuation," in *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes. Vol. 2*. eds. A. F. Healy, S. M. Kosslyn and R. M. Shiffrin (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.), 35–67.

Bjork, E. L., and Bjork, R. A. (2011). "Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning," in *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society. Vol. 2*. eds. M. A. Gernsbacher, R. W. Pew, L. M. Hough and J. R. Pomerantz (New York: Worth Publishers), 59–68.

Bjork, R. A., and Bjork, E. L. (2019). Forgetting as the friend of learning: implications for teaching and self-regulated learning. *Adv. Physiol. Educ.* 43, 164–167. doi: 10.1152/advan.00001.2019

Bjork, R. A., and Bjork, E. L. (2020). Desirable difficulties in theory and practice. *J. Appl. Res. Mem. Cogn.* 9, 475–479. doi: 10.1016/j.jarmac.2020.09.003

Bjork, E. L., Soderstrom, N. C., and Little, J. L. (2015). Can multiple-choice testing induce desirable difficulties? evidence from the laboratory and the classroom. *Am. J. Psychol.* 128, 229–239. doi: 10.5406/amerjpsyc.128.2.0229

Blake, K. R., and Gangestad, S. (2020). On attenuated interactions, measurement error, and statistical power: guidelines for social and personality psychologists. *Personal. Soc. Psychol. Bull.* 46, 1702–1711. doi: 10.1177/0146167220913363

Book, C., Byers, J., and Freeman, D. (1983). Student expectations and teacher education traditions with which we can and cannot live. *J. Teach. Educ.* 34, 9–13. doi: 10.1177/002248718303400103

Bornstein, M. H., Hahn, C. S., and Wolke, D. (2013). Systems and cascades in cognitive development and academic achievement. *Child Dev.* 84, 154–162. doi: 10.1111/j.1467-8624.2012.01849.x

Bradley, R. T., McCraty, R., Atkinson, M., Tomasino, D., Daugherty, A., and Arguelles, L. (2010). Emotion self-regulation, psychophysiological coherence, and test anxiety: results from an experiment using electrophysiological measures. *Appl. Psychophysiol. Biofeedback* 35, 261–283. doi: 10.1007/s10484-010-9134-x

Brewer, G. A., and Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *J. Mem. Lang.* 66, 407–415. doi: 10.1016/j.jml.2011.12.009

Carpenter, S. K., Lund, T. J., Coffman, C. R., Armstrong, P. I., Lamm, M. H., and Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educ. Psychol. Rev.* 28, 353–375. doi: 10.1007/s10648-015-9311-9

Cassady, J. C. (2004a). The influence of cognitive test anxiety across the learning–testing cycle. *Learn. Instr.* 14, 569–592. doi: 10.1016/j.learninstruc.2004.09.002

Cassady, J. C. (2004b). The impact of cognitive test anxiety on text comprehension and recall in the absence of external evaluative pressure. *Appl. Cogn. Psychol.* 18, 311–325. doi: 10.1002/acp.968

Cassady, J. C., and Gridley, B. E. (2005). The effects of online formative and summative assessment on test anxiety and performance. *J. Technol. Learn. Assess* 4, 4–30.

Charles, N. E., Strong, S. J., Burns, L. C., Bullerjahn, M. R., and Serafine, K. M. (2021). Increased mood disorder symptoms, perceived stress, and alcohol

use among college students during the COVID-19 pandemic. *Psychiatry Res.* 296:113706. doi: 10.1016/j.psychres.2021.113706

Chen, I., and Chang, C. C. (2009). Cognitive load theory: an empirical study of anxiety and task performance in language learning. *Electron. J. Res. Educ. Psychol.* 7, 729–746. doi: 10.25115/ejrep.v7i18.1369

Chuderski, A. (2014). High intelligence prevents the negative impact of anxiety on working memory. *Cognit. Emot.* 29, 1197–1209. doi: 10.1080/02699931.2014.969683

Daniels, L. M., Goegan, L. D., and Parker, P. C. (2021). The impact of COVID-19 triggered changes to instruction and assessment on university students' self-reported motivation, engagement and perceptions. *Soc. Psychol. Educ.* 24, 299–318. doi: 10.1007/s11218-021-09612-3

de la Fuente, J., Pachón-Basallo, M., Santos, F. H., Peralta-Sánchez, F. J., González-Torres, M. C., Artuch-Garde, R., et al. (2021). How has the COVID-19 crisis affected the academic stress of university students? The role of teachers and students. *Front. Psychol.* 12:626340. doi: 10.3389/fpsyg.2021.626340

DeLongis, A., Folkman, S., and Lazarus, R. S. (1988). The impact of daily stress on health and mood: psychological and social resources as mediators. *J. Pers. Soc. Psychol.* 54, 486–495. doi: 10.1037/0022-3514.54.3.486

DeVaney, T. A. (2010). Anxiety and attitude of graduate students in on-campus vs. online statistics courses. *J. Stat. Educ.* 18, 1–15. doi: 10.1080/10691898.2010.11889472

Diemand-Yauman, C., Oppenheimer, D. M., and Vaughan, E. B. (2011). Fortune favors the bold (and the italicized): effects of disfluency on educational outcomes. *Cognition* 118, 111–115. doi: 10.1016/j.cognition.2010.09.012

Dobson, J. L., and Linderholm, T. (2015). The effect of selected "desirable difficulties" on the ability to recall anatomy information. *Anat. Sci. Educ.* 8, 395–403. doi: 10.1002/ase.1489

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., and Willingham, D. T. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychol. Sci. Public Interest* 14, 4–58. doi: 10.1177/1529100612453266

Edwards, J. A., and Templeton, A. (2005). The structure of perceived qualities of situations. *Eur. J. Soc. Psychol.* 35, 705–723. doi: 10.1002/ejsp.271

Efklides, A., Papadaki, M., Papantoniou, G., and Kiosseoglou, G. (1997). Effects of cognitive ability and affect on school mathematics performance and feelings of difficulty. *Am. J. Psychol.* 110, 225–258. doi: 10.2307/1423716

England, B. J., Brigati, J. R., and Schussler, E. E. (2017). Student anxiety in introductory biology classrooms: perceptions about active learning and persistence in the major. *PLoS One* 12:e0182506. doi: 10.1371/journal.pone.0182506

Epel, E. S., Crosswell, A. D., Mayer, S. E., Prather, A. A., Slavich, G. M., Puterman, E., et al. (2018). More than a feeling: a unified view of stress measurement for population science. *Front. Neuroendocrinol.* 49, 146–169. doi: 10.1016/j.yfrne.2018.03.001

Eysenck, M. W., and Calvo, M. G. (1992). Anxiety and performance: the processing efficiency theory. *Cognit. Emot.* 6, 409–434. doi: 10.1080/02699939208409696

Eysenck, M. W., Derakshan, N., Santos, R., and Calvo, M. G. (2007). Anxiety and cognitive performance: attentional control theory. *Emotion* 7:336. doi: 10.1037/1528-3542.7.2.336

Faul, F., Erdfelder, E., Buchner, A., and Lang, A. G. (2009). Statistical power analyses using G* power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/BRM.41.4.1149

Feraco, T., Casali, N., Tortora, C., Dal Bon, C., Accarrino, D., and Meneghetti, C. (2020). Using Mobile devices in teaching large university classes: how does it affect exam success? *Front. Psychol.* 11:1363. doi: 10.3389/fpsyg.2020.01363

Fergus, S. (2020). Moving to online teaching–concepts, considerations and pitfalls. *LINK* 5

Fergusson, D. M., Horwood, L. J., and Ridder, E. M. (2005). Show me the child at seven II: childhood intelligence and later outcomes in adolescence and young adulthood. *J. Child Psychol. Psychiatry* 46, 850–858. doi: 10.1111/j.1469-7610.2005.01472.x

Fliege, H., Rose, M., Arck, P., Levenstein, S., and Klapp, B. F. (2001). Validierung des "perceived stress questionnaire"(PSQ) an einer deutschen Stichprobe. [validation of the "perceived stress questionnaire"(PSQ) in a German sample]. *Diagnostica* 47, 142–152. doi: 10.1026//0012-1924.47.3.142

García, T., Rodríguez, C., Betts, L., Areces, D., and González-Castro, P. (2016). How affective-motivational variables and approaches to learning relate to mathematics achievement in upper elementary levels. *Learn. Individ. Differ.* 49, 25–31. doi: 10.1016/j.lindif.2016.05.021

Goetz, T., Preckel, F., Pekrun, R., and Hall, N. C. (2007). Emotional experiences during test taking: does cognitive ability make a difference? *Learn. Individ. Differ.* 17, 3–16. doi: 10.1016/j.lindif.2006.12.002

Gottfredson, L. S. (1997). Why g matters: the complexity of everyday life. *Intelligence* 24, 79–132. doi: 10.1016/S0160-2896(97)90014-3

Greving, S., and Richter, T. (2018). Examining the testing effect in university teaching: retrievability and question format matter. *Front. Psychol.* 9:2412. doi: 10.3389/fpsyg.2018.02412

Grimaldi, P. J., and Karpicke, J. D. (2014). Guided retrieval practice of educational materials using automated scoring. *J. Educ. Psychol.* 106, 58–68. doi: 10.1037/a0033208

Hayes, A. F. (2018). *Introduction to Mediation, Moderation, and Conditional Process Analysis. 2nd Edn*. New York: The Guilford Press.

Heitmann, S., Grund, A., Berthold, K., Fries, S., and Roelle, J. (2018). Testing is more desirable when it is adaptive and still desirable when compared to note-taking. *Front. Psychol.* 9:2596. doi: 10.3389/fpsyg.2018.02596

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Rev. Educ. Res.* 58, 47–77. doi: 10.3102/00346543058001047

Hinze, S. R., and Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: performance pressure reduces the benefits of retrieval practice. *Appl. Cogn. Psychol.* 28, 597–606. doi: 10.1002/acp.3032

Hobfoll, S. E. (1989). Conservation of resources: a new attempt at conceptualizing stress. *Am. Psychol.* 44, 513–524. doi: 10.1037/0003-066X.44.3.513

Iwamoto, D. H., Hargis, J., Taitano, E. J., and Vuong, K. (2017). Analyzing the efficacy of the testing effect using KahootTM on student performance. *Turk. Online J. Dist. Educ.* 18, 80–93. doi: 10.17718/tojde.306561

Jamieson, J. P., Peters, B. J., Greenwood, E. J., and Altose, A. J. (2016). Reappraising stress arousal improves performance and reduces evaluation anxiety in classroom exam situations. *Soc. Psychol. Personal. Sci.* 7, 579–587. doi: 10.1177/1948550616644656

Jing, H. G., Szpunar, K. K., and Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *J. Exp. Psychol. Appl.* 22, 305–318. doi: 10.1037/xap0000087

Johnson, D. R., and Gronlund, S. D. (2009). Individuals lower in working memory capacity are particularly vulnerable to anxiety's disruptive effect in performance. *Anxiety Stress Coping* 22, 201–213. doi: 10.1080/10615800802291277

Kaiser, I., Mayer, J., and Malai, D. (2018). Self-generation in the context of inquiry-based learning. *Front. Psychol.* 9:2440. doi: 10.3389/fpsyg.2018.02440

Kalyuga, S., Chandler, P., Tuovinen, J., and Sweller, J. (2001). When problem solving is superior to studying worked examples. *J. Educ. Psychol.* 93, 579–588. doi: 10.1037/0022-0663.93.3.579

Karpicke, J. D. (2017). "Retrieval-based learning: A decade of progress," in *Cognitive Psychology of Memory, Vol. 2 of Learning and Memory: A Comprehensive Reference (J. H. Byrne, Series Ed.)*. ed. J. T. Wixted (Oxford: Academic Press).

Karpicke, J. D., and Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educ. Psychol. Rev.* 27, 317–326. doi: 10.1007/s10648-015-9309-3

Karpicke, J. D., Butler, A. C., and Roediger, H. L. (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own? *Memory* 17, 471–479. doi: 10.1080/09658210802647009

Kausar, R. (2010). Perceived stress, academic workloads and use of coping strategies by university students. *J. Behav. Sci.* 20, 31–45.

Keeley, J., Zayac, R., and Correia, C. (2008). Curvilinear relationships between statistics anxiety and performance among undergraduate students: evidence for optimal anxiety. *Stat. Educ. Res. J.* 7, 4–15.

Khan, M. J., Altaf, S., and Kausar, H. (2013). Effect of perceived academic stress on Students' performance. *FWU. J. Social. Sci.* 7, 146–151.

Khanna, M. M. (2015). Ungraded pop quizzes: test-enhanced learning without all the anxiety. *Teach. Psychol.* 42, 174–178. doi: 10.1177/0098628315573144

Khng, K. H. (2017). A better state-of-mind: deep breathing reduces state anxiety and enhances test performance through regulating test cognitions in children. *Cognit. Emot.* 31, 1502–1510. doi: 10.1080/02699931.2016.1233095

Kimmel, H. D., and Bevill, M. (1985). Habituation and dishabituation of the human orienting reflex under instruction-induced stress. *Physiol. Psychol.* 13, 92–94. doi: 10.3758/BF03326503

König, J., Jäger-Biela, D. J., and Glutsch, N. (2020). Adapting to online teaching during COVID-19 school closure: teacher education and teacher competence effects among early career teachers in Germany. *Eur. J. Teach. Educ.* 43, 608–622. doi: 10.1080/02619768.2020.1809650

Koriat, A., and Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *J. Mem. Lang.* 52, 478–492. doi: 10.1016/j.jml.2005.01.001

Kornell, N., Rhodes, M. G., Castel, A. D., and Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: dissociating memory, memory beliefs, and memory judgments. *Psychol. Sci.* 22, 787–794. doi: 10.1177/0956797611407929

Kuncel, N. R., Hezlett, S. A., and Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: can one construct predict them all? *J. Pers. Soc. Psychol.* 86, 148–161. doi: 10.1037/0022-3514.86.1.148

Kurebayashi, L. F. S., Do Prado, J. M., and Da Silva, M. J. P. (2012). Correlations between stress and anxiety levels in nursing students. *J. Nurs. Educ. Pract.* 2, 128. doi: 10.5430/jnep.v2n3p128

Lazarus, R. S. (1990). Theory-based stress measurement. *Psychol. Inq.* 1, 3–13. doi: 10.1207/s15327965pli0101_1

Lazarus, R. S., and Folkman, S. (1987). Transactional theory and research on emotions and coping. *Eur. J. Personal.* 1, 141–169. doi: 10.1002/per.2410010304

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teach. Psychol.* 29, 210–212. doi: 10.1207/S15328023TOP2903_06

Lehmann, J., Goussios, C., and Seufert, T. (2016). Working memory capacity and disfluency effect: an aptitude-treatment-interaction study. *Metacogn. Learn.* 11, 89–105. doi: 10.1007/s11409-015-9149-z

Leiner, J. E. M., Scherndl, T., and Ortner, T. M. (2018). How do men and women perceive a high-stakes test situation? *Front. Psychol.* 9:2216. doi: 10.3389/fpsyg.2018.02216

LePine, J. A., LePine, M. A., and Jackson, C. L. (2004). Challenge and hindrance stress: relationships with exhaustion, motivation to learn, and learning performance. *J. Appl. Psychol.* 89, 883–891. doi: 10.1037/0021-9010.89.5.883

Levenstein, S., Prantera, C., Varvo, V., Scribano, M. L., Berto, E., Luzi, C., et al. (1993). Development of the perceived stress questionnaire: a new tool for psychosomatic research. *J. Psychosom. Res.* 37, 19–32. doi: 10.1016/0022-3999(93)90120-5

Limcaoco, R. S. G., Mateos, E. M., Fernandez, J. M., and Roncero, C. (2020). Anxiety, worry and perceived stress in the world due to the COVID-19 pandemic, March 2020 preliminary results. *MedRxiv*. [Preprint]. doi: 10.1101/2020.04.03.20043992

Lipowsky, F., Richter, T., Borromeo-Ferri, R., Ebersbach, M., and Hänze, M. (2015). Wünschenswerte Erschwernisse beim Lernen. *Schulpädagogik heute* 6, 1–10.

Maass, J. K., and Pavlik, P. I. (2016). Modeling the influence of format and depth during effortful retrieval practice. in *Proceedings of the 9th International Conference on Educational Data Mining*. Berlin, Heidelberg: Springer-Verlag.

Matayoshi, J., Uzun, H., and Cosyn, E. (2020). "Studying retrieval practice in an intelligent tutoring system," in *Proceedings of the Seventh ACM Conference on Learning@ Scale*. New York: Association for Computing Machinery. 51–62. doi: 10.1145/3386527.3405927

Mavridis, A., and Tsiatsos, T. (2017). Game-based assessment: investigating the impact on test anxiety and exam performance. *J. Comput. Assist. Learn.* 33, 137–150. doi: 10.1111/jcal.12170

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., and Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: the effects of quiz frequency and placement. *J. Educ. Psychol.* 103, 399–414. doi: 10.1037/a0021782

McDaniel, M. A., Hines, R. J., and Guynn, M. J. (2002). When text difficulty benefits less-skilled readers. *J. Mem. Lang.* 46, 544–561. doi: 10.1006/jmla.2001.2819

McDaniel, M. A., Roediger, H. L., and McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychon. Bull. Rev.* 14, 200–206. doi: 10.3758/BF03194052

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., and Roediger, H. L. (2013). Quizzing in middle-school science: successful transfer performance on classroom exams. *Appl. Cogn. Psychol.* 27, 360–372. doi: 10.1002/acp.2914

McGrath, J. E. (1970). *Social and Psychological Factors in Stress*. Oxford, England: Holt, Rinehart, and Winston.

McNamara, D. S., Kintsch, E., Songer, N. B., and Kintsch, W. (1996). Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cogn. Instr.* 14, 1–43. doi: 10.1207/s1532690xci1401_1

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., and Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *J. Exp. Psychol. Learn. Mem. Cogn.* 44, 1474–1486. doi: 10.1037/xlm0000486

Mok, W. S. Y., and Chan, W. W. L. (2016). How do tests and summary writing tasks enhance long-term retention of students with different levels of test anxiety? *Instr. Sci.* 44, 567–581. doi: 10.1007/s11251-016-9393-x

Muthuprasad, T., Aiswarya, S., Aditya, K. S., and Jha, G. K. (2021). Students' perception and preference for online education in India during COVID-19 pandemic. *Social. Sci. Humanities. Open.* 3:100101. doi: 10.1016/j.ssaho.2020.100101

Naveh-Benjamin, M. (1991). A comparison of training programs intended for different types of test-anxious students: further support for an information-processing model. *J. Educ. Psychol.* 83, 134–139. doi: 10.1037/0022-0663.83.1.134

Nieto-Escamez, F. A., and Roldán-Tapia, M. D. (2021). Gamification as online teaching strategy during COVID-19: a mini-review. *Front. Psychol.* 12:648552. doi: 10.3389/fpsyg.2021.648552

Nouhan, C., Scott, N., and Womack, J. (2021). Emergent role of artificial intelligence in higher education. *IEEE Future Directions Newsl. Technol. Policy Ethics* 31

Nyroos, M., Schéle, I., and Wiklund-Hörnqvist, C. (2016). Implementing test enhanced learning: Swedish teacher students' perception of quizzing. *Int. J. Higher. Educ.* 5, 1–12. doi: 10.5430/ijhe.v5n4p1

Oberauer, K., Schulze, R., Wilhelm, O., and Süß, H. M. (2005). Working memory and intelligence—their correlation and their relation: comment on Ackerman, Beier, and Boyle (2005). *Psychol. Bull.* 131, 61–65. doi: 10.1037/0033-2909.131.1.61

Obrero-Gaitán, E., Nieto-Escamez, F., Zagalaz-Anula, N., and Cortés-Pérez, I. (2021). An innovative approach for online neuroanatomy and neuropathology teaching based on 3D virtual anatomical models using leap motion controller during COVID-19 pandemic. *Front. Psychol.* 12:1853. doi: 10.3389/fpsyg.2021.590196

O'Neil, J. H., Spielberger, C. D., and Hansen, D. N. (1969). Effects of state anxiety and task difficulty on computer-assisted learning. *J. Educ. Psychol.* 60, 343–350. doi: 10.1037/h0028323

Oppenheimer, D. M., and Alter, A. L. (2014). The search for moderators in disfluency research. *Appl. Cogn. Psychol.* 28, 502–504. doi: 10.1002/acp.3023

Orlov, G., McKee, D., Berry, J., Boyle, A., DiCiccio, T., Ransom, T., et al. (2021). Learning during the COVID-19 pandemic: it is not who you teach, but how you teach. *Econ. Lett.* 202:109812. doi: 10.1016/j.econlet.2021.109812

Owens, M., Stevenson, J., Hadwin, J. A., and Norgate, R. (2014). When does anxiety help or hinder cognitive test performance? the role of working memory capacity. *Br. J. Psychol.* 105, 92–101. doi: 10.1111/bjop.12009

Park, J., and Choi, B. C. (2008). Higher retention after a new take-home computerised test. *Br. J. Educ. Technol.* 39, 538–547. doi: 10.1111/j.1467-8535.2007.00752.x

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., et al (2007). *Organizing Instruction and Study to Improve Student Learning. IES Practice Guide. NCER 2007--2004*. Washington, DC: National Center for Education Research.

Pinel, P. J., and Pauli, P. (2012). *Biopsychologie [Biopsychology]*. 8. Auflage Edn. München: Pearson Education.

Pozo, J. I., Echeverría, M. P. P., Cabellos, B., and Sánchez, D. L. (2021). Teaching and learning in times of COVID-19: uses of digital technologies during school lockdowns. *Front. Psychol.* 12:656776. doi: 10.3389/fpsyg.2021.656776

Pugh, D., De Champlain, A., Gierl, M., Lai, H., and Touchie, C. (2020). Can automated item generation be used to develop high quality MCQs that assess application of knowledge? *Res. Pract. Technol. Enhanc. Learn.* 15, 1–13. doi: 10.1186/s41039-020-00134-8

Pyc, M. A., and Rawson, K. A. (2009). Testing the retrieval effort hypothesis: does greater difficulty correctly recalling information lead to higher levels of memory? *J. Mem. Lang.* 60, 437–447. doi: 10.1016/j.jml.2009.01.004

Radanliev, P., De Roure, D., Van Kleek, M., Santos, O., and Ani, U. (2020). Artificial intelligence in cyber physical systems. *AI. Soc.*, 1–14. doi: 10.1007/s00146-020-01049-0, [Epub ahead of print].

Reeve, C. L., Bonaccio, S., and Winford, E. C. (2014). Cognitive ability, exam-related emotions and exam performance: a field study in a college setting. *Contemp. Educ. Psychol.* 39, 124–133. doi: 10.1016/j.cedpsych.2014.03.001

Richland, L. E., Bjork, R. A., Finley, J. R., and Linn, M. C. (2005). "Linking cognitive science to education: generation and interleaving effects," in *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*. eds. B. G. Bara, L. Barsalou and M. Bucciarelli (Mahwah, NJ: Erlbaum), 1850–1855.

Robey, A. M. (2017). The benefits of testing: individual differences based on student factors. *J. Mem. Lang.* 108:104029. doi: 10.1016/j.jml.2019.104029

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., and McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *J. Exp. Psychol. Appl.* 17, 382–395. doi: 10.1037/a0026252

Roediger, H. L., and Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychol. Sci.* 17, 249–255. doi: 10.1111/j.1467-9280.2006.01693.x

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., and Spinath, F. M. (2015). Intelligence and school grades: a meta-analysis. *Intelligence* 53, 118–137. doi: 10.1016/j.intell.2015.09.002

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychol. Bull.* 140, 1432–1463. doi: 10.1037/a0037559

Rummer, R., Schweppe, J., Gerst, K., and Wagner, S. (2017). Is testing a more effective learning strategy than note-taking? *J. Exp. Psychol. Appl.* 23, 293–300. doi: 10.1037/xap0000134

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: reactions to tests. *J. Pers. Soc. Psychol.* 46:929. doi: 10.1037/0022-3514.46.4.929

Sarason, I. G., and Sarason, B. R. (1990). "Test anxiety," in *Handbook of Social and Evaluation Anxiety*. ed. H. Leitenberg (Boston, MA: Springer)

Sarju, J. P. (2020). Rapid adaptation of a traditional introductory lecture course on catalysis into content for remote delivery online in response to global pandemic. *J. Chem. Educ.* 97, 2590–2597. doi: 10.1021/acs.jchemed.0c00786

Schmohl, T., Schwickert, S., and Glahn, O. (2020). "Conceptual design of an AI-based learning assistant," in *Filodiritto Editore – 10th International Conference The Future of Education – Virtual Edition. Vol. 10*; June 18–19, 2020; (Florence, Italy: Filodiritto), 309–313.

Schunk, H. D., and Gaa, J. P. (1981). Goal-setting influence on learning and self-evaluation. *J. Classroom. Interact.* 16, 38–44.

Seipp, B. (1991). Anxiety and academic performance: a meta-analysis of findings. *Anxiety Res.* 4, 27–41. doi: 10.1080/08917779108248762

Sotardi, V. A., Bosch, J., and Brogt, E. (2020). Multidimensional influences of anxiety and assessment type on task performance. *Soc. Psychol. Educ.* 23, 499–522. doi: 10.1007/s11218-019-09508-3

Stadler, M., Becker, N., Gödker, M., Leutner, D., and Greiff, S. (2015). Complex problem solving and intelligence: a meta-analysis. *Intelligence* 53, 92–101. doi: 10.1016/j.intell.2015.09.005

Stern, E. (2015). "Intelligence, prior knowledge, and learning," in *International Encyclopedia of the Social and Behavioral Sciences. 2nd Edn. Vol. 12* (Oxford, United Kingdom: Elsevier), 323–328.

Stern, E. (2017). Individual differences in the learning potential of human beings. *npj. Sci. Learn.* 2:2. doi: 10.1038/s41539-016-0003-0

Sternberg, R. J. (1997). The concept of intelligence and its role in lifelong learning and success. *Am. Psychol.* 52, 1030–1037. doi: 10.1037/0003-066X.52.10.1030

Strenze, T. (2015). Intelligence and socioeconomic success: a study of correlations, causes and consequences. doctoral dissertation. Tartu: Tartu University.

Struthers, C. W., Perry, R. P., and Menec, V. H. (2000). An examination of the relationship among academic stress, coping, motivation, and performance in college. *Res. High. Educ.* 41, 581–592. doi: 10.1023/A:1007094931292

Sung, Y. T., Chao, T. Y., and Tseng, F. L. (2016). Reexamining the relationship between test anxiety and learning achievement: an individual-differences perspective. *Contemp. Educ. Psychol.* 46, 241–252. doi: 10.1016/j.cedpsych.2016.07.001

Tobias, S. (1984). "Test Anxiety: Cognitive Interference or Inadequate Preparation?" in *Annual Meeting of the American Educational Research Association*. April 23–27, 1984; New Orleans, LA.

Tse, C. S., and Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *J. Exp. Psychol. Appl.* 18, 253–264. doi: 10.1037/a0029190

Tyler, S. W., Hertel, P. T., McCallum, M. C., and Ellis, H. C. (1979). Cognitive effort and memory. *J. Exp. Psychol. Hum. Learn. Mem.* 5, 607–617. doi: 10.1037/0278-7393.5.6.607

Wang, T., Ren, X., and Schweizer, K. (2017). Learning and retrieval processes predict fluid intelligence over and above working memory. *Intelligence* 61, 29–36. doi: 10.1016/j.intell.2016.12.005

Wang, A. I., and Tahir, R. (2020). The effect of using Kahoot! for learning–a literature review. *Comput. Educ.* 149:103818. doi: 10.1016/j.compedu.2020.103818

Wenzel, K., and Reinhard, M.-A. (2019). Relatively unintelligent individuals do not benefit from intentionally hindered learning: the role of desirable difficulties. *Intelligence* 77:101405. doi: 10.1016/j.intell.2019.101405

Wenzel, K., and Reinhard, M.-A. (2021). Does the end justify the means? learning tests lead to more negative evaluations and to more stress experiences. *Learn. Motiv.* 73:101706. doi: 10.1016/j.lmot.2020.101706

Wu, S., Zhang, K., Parks-Stamm, E. J., Hu, Z., Ji, Y., and Cui, X. (2021). Increases in anxiety and depression during COVID-19: a large longitudinal study from China. *Front. Psychol.* 12:706601. doi: 10.3389/fpsyg.2021.706601

Yang, C., Chen, A., and Chen, Y. (2021). College students' stress and health in the COVID-19 pandemic: the role of academic workload, separation from school, and fears of contagion. *PLoS One* 16:e0246676. doi: 10.1371/journal.pone.0246676

Yang, C., Luo, L., Vadillo, M. A., Yu, R., and Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: a systematic and meta-analytic review. *Psychol. Bull.* 147, 399–435. doi: 10.1037/bul0000309

Yu, H., Liu, P., Huang, X., and Cao, Y. (2021). Teacher online informal learning as a means to innovative teaching during home quarantine in the COVID-19 pandemic. *Front. Psychol.* 12:596582. doi: 10.3389/fpsyg.2021.596582

Zeidner, M. (1998). "Perspectives on Individual Differences," in *Test Anxiety: The State of the Art* (New York: Plenum Press)

Zhang, Y., and Liu, B. (2021). Psychological distress among Chinese college students during the COVID-19 pandemic: does attitude toward online courses matter? *Front. Psychol.* 12:665525. doi: 10.3389/fpsyg.2021.665525

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX A – MATERIALS

Materials and example items (translated for this presentation, used materials in German)

Example questions of the questions applied in the learning test in the test condition:

1. What is meant by cerebral dominance?

Please answer the question in one or two sentences at most.

2. In apraxia, which type of motor function/movement is disturbed: _____
3. What should the patient enumerate during the sodium amytal test?

(a) Nothing
(b) Difficult things (e.g., answers to complex math problems, statements,…)
(c) Well-known things (e.g., the letters of the alphabet, the days of the week,…)
(d) Made-up things (e.g., freely invented names,…)

Manipulation check questions applied at the end of Session 1:

1. How difficult did you find working on the second learning phase? one (*very easy*) to five (*very difficult*).
2. How helpful for retaining the learning material did you find working on the second learning phase? one (*not helpful at all*) to five (*very helpful*).
3. How (cognitively) strenuous did you find working on the second learning phase? one (*not strenuous at all*) to five (*very strenuous*).
4. How would you most likely evaluate the second learning phase? As …, one (*a challenge*) to five (*a threat*).
5. How would you best describe the second learning phase? As…, one (*extremely negative*) to five (*extremely positive*).
6. How well do you think you have worked through the second learning phase? one (*very poor*) to five (*very well*).

APPENDIX B – EXPLORATORY ANALYSES

Exploratory analyses focusing on the three different types of final test questions:

Considering only the final test questions that were identical to the questions posed in the learning test in Session 1 (following: *identical final test questions*), participants were on average able to give 6.51 of 11 (59.18%) correct answers ($SD = 2.08$, range: 1–10). We then conducted a *t*-test to compare later learning outcomes indicated only by the identical final test questions for participants in both learning conditions: $M_{\text{re-reading}} = 6.00$, $SD_{\text{re-reading}} = 2.16$, $M_{\text{test}} = 7.07$, $SD_{\text{test}} = 1.87$, $t(87) = -2.49$, $p = 0.015$, $d = -0.53$ (95% CI[−0.95; −0.10]). As assumed, participants in the test condition answered more identical final test questions correctly than participants in the re-reading control condition. The size of this effect can be interpreted as medium.

Considering only the final test questions that were slightly changed versions of questions posed in the learning test in Session 1 to assess transfer (following: *transfer final test questions*), participants were on average able to give 3.41 of 9 (37.89%) correct answers ($SD = 1.80$, range: 0–8). We then conducted a *t*-test to compare later learning outcomes indicated only by the transfer final test questions for participants in both learning conditions: $M_{\text{re-reading}} = 3.02$, $SD_{\text{re-reading}} = 1.60$, $M_{\text{test}} = 3.83$, $SD_{\text{test}} = 1.92$, $t(87) = -2.18$, $p = 0.032$, $d = -0.46$ (95% CI[−0.88; −0.04]). As assumed, participants in the test condition answered more transfer final test questions correctly than participants in the re-reading control condition. The size of this effect can be interpreted as medium.

Considering only the final test questions that were new and focused on information that were presented in the textbook chapter but that had not been implemented in the learning test in Session 1 (following: *new final test questions*), participants were on average able to give 3.93 of 7 (56.14%) correct answers ($SD = 4.43$, range: 1–7). We then conducted a *t*-test to compare later learning outcomes indicated only by the new final test questions for participants in both learning conditions: $M_{\text{re-reading}} = 3.85$, $SD_{\text{re-reading}} = 1.33$, $M_{\text{test}} = 4.03$, $SD_{\text{test}} = 1.54$, $t(87) = -0.57$, $p = 0.572$, $d = -0.12$ (95% CI[−0.54; 0.27]). Participants in the test condition did not significantly answer more new final test questions correctly than participants in the re-reading control condition.

Notably, these explorative findings indicate that the beneficial effects of tests only arise for information that were actually worked on during the learning test and not for information that participants read in the initial study opportunity but that had not been part of the learning test.

Exploratory correlational analyses showed that participants stress perceptions were negatively correlated to identical final test questions ($r = -0.18$, $p = 0.095$), transfer final test questions ($r = -0.26$, $p = 0.014$; showing a small to medium correlation), and new final test questions ($r = -0.20$, $p = 0.055$). Notably, only the correlation of transfer final test questions and participants stress perception reached significance when using two-sided tests (the correlations among stress perceptions and identical as well as new final test questions can be described as marginally significant and reached significance when using one-sided tests). Further exploratory analyses yielded that the three correlation coefficients did not significantly differ from each other (all $ps \geq 0.232$).

Further exploratory correlational analyses found that participants' intelligence-estimates were significantly and positively correlated to identical final test questions ($r = 0.28$, $p = 0.009$), transfer final test questions ($r = 0.30$, $p = 0.005$), and new final test questions ($r = 0.25$, $p = 0.017$). Thus, higher intelligence-estimates were generally linked to higher later learning outcomes for the three different types of final test questions (showing medium correlations). Further exploratory analyses showed that the three correlation coefficients did not significantly differ from each other (all $ps \geq 0.321$).

Exploratory analyses focusing on the correlations among participant stress perceptions and the six questions checking the manipulation of the two learning conditions:

TABLE 1 | Exploratory correlations among the six manipulation check questions and participants stress perceptions ($N = 89$).

	1	2	3	4	5	6	7
1. Difficulty	—						
2. Helpfulness	–0.17	—					
3. Strenuousness	0.59**	0.14	—				
4. Evaluation – challenge/threat	–0.13	–0.21*	–0.18	—			
5. Evaluation – negative/positive	–0.43**	0.60**	–0.11	–0.05	—		
6. Successfulness	–0.66**	0.35**	–0.36**	0.09	0.51**	—	
7. Stress perceptions	0.51**	–0.28**	0.31**	0.04	–0.44**	–0.67**	—

*$p < 0.05$; **$p < 0.01$ See **Appendix A** *for a full list of the manipulation check questions.*

ACKNOWLEDGMENTS / DANKSAGUNG

Zuallererst gilt mein Dank natürlich meinem Betreuer Marc, ohne welchen diese Doktorarbeit erst gar nicht möglich gewesen wäre. Vielen Dank, dass du mir damals eine Chance gegeben und mich danach immer weiter angetrieben hast. Und so sehr ich es eigentlich nie öffentlich zugegeben wollte, hat sich letztendlich „*Paper First*“ doch als ganz gute Strategie herausgestellt.

Als nächstes möchte ich bei der Universität Kassel für das erhaltene Promotionsstipendium bedanken, welches es mir ermöglichte, mich ganz dieser Doktorarbeit zu widmen. Vielen Dank an die Graduiertenförderung der Universität Kassel für das Fördern meiner Promotion.

Außerdem bin ich allen Personen dankbar, welche an der Entstehung meiner publizierten Artikel beteiligt waren—dies beinhaltet besonders meine Ko-Autor*innen, die studentischen Hilfskräfte und letztendlich auch alle Versuchspersonen, welche an meinen Studien teilgenommen haben.

Weiterhin möchte ich gerne meinen Gutachter*innen und der Kommission dafür danken, dass sie sich die Zeit nehmen, um diese Doktorarbeit zu lesen, zu bewerten und mit mir darüber zu diskutieren—ich hoffe sie finden das Ganze genauso spannend und wichtig wie ich.

Passend dazu möchte ich mich auch generell noch einmal bei Mirjam dafür bedanken, dass ich beinahe mein gesamtes Studium als studentische Hilfskraft in der Entwicklungspsychologie verbringen und dort erste Forschungsluft schnuppern durfte. Genauso möchte ich mich bei Ralf für die herzliche Aufnahme in der Allgemeinen Psychologie und für die Möglichkeit dort zu arbeiten und weiter zu forschen bedanken.

Eigentlich könnte ich an dieser Stelle (und werde es auch einfach tun) noch das gesamte Institut für Psychologie erwähnen und mich dort bei allen Mitarbeiter*innen bedanken mit welchen ich dort zu tun hatte—ich habe mich während meiner gesamten Promotionszeit im Institut immer sehr willkommen und unterstützt gefühlt.

Genauso gilt mein Dank meinen Kolleg*innen (bei denen es sich wunderbarerweise mittlerweile um Freund*innen handelt). Dies beinhaltet beispielsweise alle (jetzigen und ehemaligen) Mitarbeiter*innen der Sozialpsychologie, mit welchen das gemeinsame Arbeiten und die Doktorandinnen-Treffen immer eine Freude waren, sowie die Mitarbeiterinnen in der Allgemeinen Psychologie, zu welchen ich seit kurzer Zeit auch zählen darf. Betonen möchte ich in diesem Rahmen aber auch die Mitglieder der sogenannten *„Mensa-Gruppe“*, mit welchen ich viele schöne Stunden verbringen durfte—sei es beim Mittagessen, Kaffeetrinken oder dem ein oder anderen privaten Treffen. Die vielen Gespräche und gemeinsamen Aktivitäten während meiner Promotionszeit würde ich auf keinen Fall missen wollen (dazu zählt unter anderem auch das gemeinsame Bücher-Vorlesen im Leo). Ganz lieben Dank für eure durchgängige Unterstützung und Freundschaft.

Besonders möchte ich mich am Ende natürlich auch bei meinem Freundeskreis und meiner Familie bedanken, ohne welche dies wirklich gar nicht möglich gewesen wäre. Lieben Dank dabei beispielsweise an Bella und Alessa (und natürlich an Jessi, welcher ich das heutige Endprodukt gerne gezeigt hätte). Ganz herzlicher Dank geht in diesem Rahmen aber natürlich an meine Eltern und meine Schwester (und an die Tiere). Danke, dass ihr mich angespornt habt durchzuhalten, wenn ich dachte, dass ich niemals fertig werden würde. Danke, dass ihr mich häufig gezwungen habt schöne Sachen zu unternehmen und meinen Laptop zu ignorieren. Und Danke, dass ihr euch so oft und so entspannt (bzw. an den passenden Stellen entsetzt oder entrüstet) mein Gemecker und Gejammer angehört habt. Entschuldigt noch einmal, dass ich in letzter Zeit von eigentlich nichts anderem außer dieser

Doktorarbeit reden konnte und eigentlich zu nichts wirklich zu gebrauchen war—hoffen wir mal, dass das jetzt wieder besser wird.