Computers in Human Behavior Reports 15 (2024) 100454

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

The relation between learners' experience in simulations and diagnostic accuracy: Generalizability across medical and teacher education

Olga Chernikova [a,*], Matthias Stadler [e], Daniel Sommerhoff [b], Christian Schons [c], Nicole Heitzmann [a], Doris Holzberger [c], Tina Seidel [c], Constanze Richters [a], Amadeus J. Pickal [d], Christof Wecker [d], Michael Nickl [b,c], Elias Codreanu [c], Stefan Ufer [a], Stephanie Kron [a], Caroline Corves [e], Birgit J. Neuhaus [a], Martin R. Fischer [e], Frank Fischer [a]

[a] University of Munich (LMU), Germany
[b] Leibniz Institute for Science and Mathematics Education (IPN), University of Kiel (CAU), Germany
[c] Technical University of Munich (TUM), School of Social Sciences and Technology, Germany
[d] University of Hildesheim, Germany
[e] Institute for Medical Education at the University Hospital, LMU Munich, Germany

ARTICLE INFO

Keywords:
Simulation-based learning
Diagnostic skills
Learners' experience
Higher education
Cross-domain research

ABSTRACT

Simulation-based learning is being increasingly implemented across different domains of higher education to facilitate essential skills and competences (e.g. diagnostic skills, problem-solving, etc.). However, the lack of research that assesses and compares simulations used in different contexts (e.g., from design perspective) makes it challenging to effectively transfer good practices or establish guidelines for effective simulations across different domains. This study suggests some initial steps to address this issue by investigating the relations between learners' experience in simulation-based learning environments and learners' diagnostic accuracy across several different domains and types of simulations, with the goal of facilitating cross-domain research and generalizability. The findings demonstrate that used learners' experience ratings are correlated with objective performance measures, and can be used for meaningful comparisons across different domains. Measures of perceived extraneous cognitive load were found to be specific to the simulation and situation, while perceived involvement and authenticity were not. Further, the negative correlation between perceived extraneous cognitive load and perceived authenticity was more pronounced in interaction-based simulations. These results provide supporting evidence for theoretical models that highlight the connection between learners' experience in simulated learning environments and their performance. Overall, this research contributes to the understanding of the relationship between learners' experience in simulation-based learning environments and their diagnostic accuracy, paving the way for the dissemination of best practices across different domains within higher education.

1. Problem statement

As learners in higher education need to be prepared for their future profession, their professional competences should include a range of complex skills. The most critical skills across different higher education subjects are analytic and diagnostic skills that enable the learner (and a professional) to collect, evaluate, and utilize relevant information to arrive at and communicate professional decisions (e.g., Gartmeier et al., 2015).

Simulation-based learning environments are increasingly employed in higher education to assess and cultivate diagnostic skills (e.g., Bygstad et al., 2022; Cook, 2014; Chernikova et al., 2020). Different types of simulated situations allow addressing a whole range of professional skills in higher education in safe (e.g., Ziv et al., 2003) and engaging learning environments with simulations differing in the context domain or the source of diagnostic information (Heitzmann et al., 2019). To provide realistic training possibilities, simulation-based learning environments should balance authenticity, involvement, and cognitive demand (e.g., Brom et al., 2017; Seidel et al., 2010). Accordingly, research on simulation-based learning increasingly investigates learners'

* Corresponding author. R1112, Leopoldstr. 13, 80802, Munich, Germany.
E-mail address: o.chernikova@psy.lmu.de (O. Chernikova).

https://doi.org/10.1016/j.chbr.2024.100454
Received 9 May 2024; Received in revised form 24 June 2024; Accepted 3 July 2024
Available online 5 July 2024
2451-9588/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

experience of simulations and its importance for learning outcomes (e.g., Codreanu et al., 2020; Gegenfurtner et al., 2014; Lesā et al., 2021). However, it is unclear whether the findings related to learners' experience of a simulation can be generalized across domains in the field of learning and instruction (e.g., Heitzmann et al., 2021). Addressing this problem can only take place if the generalizability of the learners' experience measures across different contexts is granted (Yarkoni, 2020). Research programs aiming at replicability and generalization of findings on learners' experience of simulations across domains (e.g., Fink et al., 2021) have rarely systematically investigated the extent to which measures with assumed cross-domain validity are actually invariant across domains. In other words, some cross-domain research papers might assume that the authenticity of, for instance, diagnosing a simulated patient in medical education can be rated on the same scale as the authenticity of another diagnostic simulation, such as a simulated interaction with a student in teacher education, and the values of this scale have the same meaning in both situations (e.g., Chernikova et al., 2020). However, this assumption may be problematic, as it may lead to an underestimation of the possible differences in understanding different phenomena across domains. Teachers and physicians might have different perspectives on the factors that contribute to authenticity.

To enable systematic development of effective simulations and transfer of good practices across domains, it is important to assess and compare simulations used in different contexts. To do so, it is essential to first assess the invariance of measurement tools used (e.g., Yarkoni, 2020) to assess learners' experience of simulations. Although non-invariance across groups does not necessarily limit the power of cross-group comparisons (e.g., Robitzsch & Lüdtke, 2020), invariance of measurement is a prerequisite for the generalizability and interpretation of results. Furthermore, from a theory perspective, understanding how learning with simulations works is crucial, as the fundamental mechanisms are likely to be consistent across contexts.

In this study, we investigate whether established subjective scales for estimating learners' experience of simulated learning environments (perceived extraneous cognitive load, perceived authenticity, and perceived involvement) show measurement invariance across different domains and types of simulations. Further, we explore the relations between learners' experience within simulation-based learning environments and learners'diagnostic accuracy.

2. Theoretical background

2.1. Diagnostic competences and diagnostic accuracy as a measure of performance in higher education

Making professional decisions is hardly possible without a thorough analysis, the ability to identify, understand, and predict events or situations by applying domain-specific knowledge and experience as well as a range of complex skills (e.g., analytic, reasoning, and problem-solving skills). Taken together, these processes form the core of diagnostic competencies (Charlin et al., 2000; Helmke et al., 2012; Spinath, 2005). There are at least two prominent domains that share diagnostic processes and activities as well as underlying competences required to come to decisions. Despite teacher and medical education domains have different professional contexts and relevant situations; there is a range of similarities (e.g., Gartmeier et al., 2015). Although in medical context diagnosing is often directed towards identifying a disease, there are also comparable activities in teacher everyday practice, for example identifying a misconception in math, identifying the competence level the student is at, etc. In teacher education diagnosing is rather related to the concepts of professional vision and formative assessment (e.g, Seidel et al., 2010).

Studies conducted in higher education often employ simulation-based learning to facilitate diagnostic competences, utilizing various indicators to measure performance and learning outcomes (e.g., (Fischer et al., 2022)). One commonly used performance indicator is judgment or diagnostic accuracy, which assesses the alignment between teachers' assessment of students' understanding and the students' actual understanding, as evaluated by an independent test or expert rating, particularly in the context of teacher education (e.g., Artelt & Rausch, 2014; Südkamp et al., 2012). In medical education, diagnostic accuracy refers to the level of agreement between the diagnosis provided and the actual presence of a disease or the correct solution for a given sample (e.g., Hege et al., 2018).

2.2. Simulation-based learning environments

There are many different operationalizations of simulations in different research fields. Simulations can be restricted to mathematically quantified modeling of relations and processes to understand particular phenomenon (e.g., in physics), or they can be applied in a much broader context. In discovery and inquiry learning, simulations are used to represent a certain aspect of reality. A simulation can be understood as a model of a real system that makes it possible to obtain knowledge about the relations among the variables of a system or to test and practice strategies for influencing and controlling a system (Frasson & Blanchard, 2012). Within the simulation-based learning environment, a comprehensive understanding of variables and their relationships is attained through the manipulation and measurement of these variables. For instance, in the field of chemical reactions, studies such as Reeves et al. (2021) emphasize the importance of this knowledge acquisition.

In the context of training (e.g., in teacher or medical education), simulations are utilized as action scenarios that require the acquisition of complex competences. This involves applying conceptual knowledge, making informed decisions, and taking appropriate actions to modify the simulated environment (Al-Kadi & Donnon, 2013).

In this study, we view simulations as approximations of practice (Grossman et al., 2009), as they enable modifying and reducing the complexity of authentic situations (decomposition) but also represent the complex nature of actual practice (approximation). Thus, a central aim of simulations is to create learning opportunities that accurately represent specific real-life problem scenarios, enabling the development of knowledge and skills and allowing practice in performing specific actions within complex situations. They are designed to emulate real-life conditions to the extent necessary for effective learning while also offering a controlled environment in which disruptive factors can be managed.

To sum up, in this paper we focus on the context of simulation-based learning environments in which (1) aspects of reality (e.g., pupil or patient behavior) are emulated in a way appropriate to the intended purpose (e.g., providing sufficient information to make a diagnostic decision), and (2) learners (within simulation-based environment) can intervene in the situation such that the further course of events depends, at least to some extent, on their actions.

Empirical research provides support for the effectiveness of learning with simulations in different domains of higher education (e.g., Cook, 2014; Theelen et al., 2019). The meta-analysis by Chernikova et al. (2020) suggests that there are multiple factors that affect the learning benefits of simulation-based learning, including the type of simulation. This study adopts the classification of simulation types based on the source of information developed by Heitzmann et al. (2019). *Document-based simulations* incorporate scenarios in which the relevant information is available as text (e.g., students' homework or patients' history), images (e.g., a drawing or x-ray), or video (e.g., recording of the lesson or interview). In contrast, *interaction-based simulations* incorporate scenarios where the relevant information is collected through interaction with the patient or student (history-taking, interviewing).

Prior research has shown that the type of simulation explains more of the variance in the effects of simulated learning environments on learning outcomes than differences in learning domains (e.g., Authors, 20XX). This may be due to differences in the cognitive demand of particular simulation types as well as differences in learners' experience

of simulated learning environments (e.g., Gegenfurtner et al., 2014). For example, a simulated situation can involve interaction with standardized patients, which requires an immediate reaction to the patients' answers and little time to decide on the next question. However, it also can involve working with written medical records, where more time is available to make decisions and it is possible to review particular aspects and check them again or consult with a book or peer. Both situations can be perceived as highly authentic, depending on the task trained, but the cognitive demands are obviously different.

2.3. Learners' experience of simulation-based learning environments

Establishing a strong connection between the content taught in higher education institutions and real-world professional situations can support future professionals in acquiring practice-oriented professional knowledge, skills, and competences (e.g., Blömeke et al., 2015). To facilitate this connection, it is essential to employ effective instructional design principles in the creation of learning environments. Such designs should consider the desired learning objectives as well as the learners' experience, ensuring a meaningful and impactful educational experience.

To obtain the most benefits from learning environments, learners should focus their attention on the learning situation (Witmer & Singer, 1998), which is known as *presence* in the simulated learning environment. Presence refers to the degree to which users of a virtual environment feel involved with, absorbed in, and engrossed by stimuli from the virtual environment (Palmer, 1995; for an overview of definitions see Agrewal et al., 2020). Conceptually, presence consists of two core dimensions: First, learners feel that they are physically (and cognitively) involved in the environment presented in the simulation, and second, the interaction with the learning environment is experienced as ecologically valid/authentic (Ijsselsteijn & Riva, 2003; Wirth et al., 2007).

Research has shown that learners benefit from simulation-based learning environments they perceive as authentic (e.g., Seidel et al., 2010) and cognitively involving (e.g., Dankbaar et al., 2016; Schubert et al., 2001). Therefore, we hypothesize that higher levels of perceived authenticity and involvement would be related to better performance (e.g., higher diagnostic accuracy) in simulation-based learning environments.

Both the availability of relevant information and the effort required to deal with the features of the learning environment (e.g., for navigation) that are not directly related to the learning task, can be conceptualized as extraneous cognitive load (e.g., Eysink et al., 2009), which in turn can also influence the experience of simulated situations and performance. We assume that higher levels of extraneous load might hinder learning and be negatively related to performance measures and learning gains. In case of this study it means, decreased diagnostic accuracy might be associated with increased levels of perceived extraneous cognitive load.

Importantly, learners may differ in how they evaluate their experience of a simulated learning environment, based on their experience, expectations and individual differences. For instance, adult learners have encountered various medical and classroom interactions in their roles as patients and students, respectively, as well as through different media sources (e.g., books, movies, newspapers). Even novice learners possess a basic understanding of what constitutes an ecologically valid interaction, although their experiences as physicians or teachers may be limited. Due to their limited experience, novices may differ from experts not only in the quantity but also in the quality of their evaluations. They may prioritize different aspects of the simulated learning environment, such authenticity, engagement, or cognitive demand based on their unique perspectives.

Similarly, teachers may have different evaluations compared to physicians when experiencing simulated learning environments. While working with medical or study record may appear similar, the contexts in which physician–patient and teacher–student interaction occur differs significantly. For example, teachers often have less one-on-one communication with individual students and may perceive such interactions as less authentic compared to interactions with group of students. Additionally, teachers, in general, may have less experience with simulation-based learning environments compared to physicians, leading them to establish different reference points based on their own experiences (e.g., give more weight to different items of the scale). Such differences might lead to difficulties (and inconsistencies) comparing learners' experiences in simulation-based learning environments across contexts and need to be addressed.

2.4. This study

The study aims to provide initial insights into design of effective simulations by examining the connection between learners' experiences within complex simulated learning environments and their performance across different types of simulations and domains. This can only be performed if the generalizability of the learners' experience measures is granted (Yarkoni, 2020). More specifically, we will focus on simulated learning environments developed to assess and foster diagnostic competence, a highly complex and relevant set of skills in various domains. Our investigation will encompass both document-based and interaction-based simulation in the contexts of both medical and teacher education.

We assume that learners' perceived authenticity and involvement are positively related to diagnostic accuracy, whereas higher perceived extraneous cognitive load is negatively related to diagnostic accuracy. We investigate these assumptions using the following research questions.

1. What is the relation between learners' experience measures (perceived measures of authenticity, involvement, and extraneous cognitive load), and how do they relate to diagnostic accuracy?
2. Do these relations differ significantly between domains and types of simulations (based on sources of relevant information)?

3. Method

3.1. Samples and procedure

The data to investigate the proposed research questions were collected through seven different studies, details of which are presented in Table 1. Each study was approved by the ethics committee of the respective host university.

To distinguish between simulation types, this study classifies diagnostic situations as document- or interaction-based simulations (Heitzmann et al., 2019). Simulations were coded as *document-based* if the diagnosis was based on information retrieved from a written document or video material and as *interaction-based* if the diagnosis was based on information obtained through interaction with the patient/pupil via an interview or similar procedure. Table 1 includes this classification as well as additional contextual information about the included studies (domain, sample, and diagnostic task). A more comprehensive description of the simulations used in this study as well as the implemented interventions and prompts can be found in the respective chapters of the Learning to Diagnose with Simulations book (Fischer & Opitz, 2022).

3.2. Measures

Participants' experience in simulations was measured using the three scales described below, which assess participants' presence based on their feelings of involvement and authenticity as well as cognitive demand (extraneous cognitive load). *Diagnostic accuracy* was operationalized as a performance measure. Most primary studies used different entry methods for making the final diagnosis (e.g., free-text entry; single

Table 1
Studies included in the analysis.

Study	N	Simulation type	Domain	Sample	Diagnostic task
1	91	document-based (written material)	Teacher Education: Mathematics	pre-service teachers Gender: 77 f/14 m Age: 22.88 (2.97) Semester: 3.86 (1.45)	diagnosing primary students' mathematical competence levels and misconceptions
2	28	document-based (video vignettes)	Teacher Education: Mathematics	pre-service teachers Gender: 13 f/15 m Age: 24 (2.97) Semester: 3.86 (2.08)	diagnosing mathematical argumentation skills
3	15	document-based (video vignettes)	Teacher Education: Biology	in-service teachers Gender: 11 f/3 m/1na Age: 24 (2.97) Semester: n/a	diagnosing the instructional quality of biology lessons
4	86	interaction-based (video/live interviews)	Medical Education	pre-service medical students Gender: 54 f/32 m Age: 26.03 (4.71) Semester: 47 in 1st sem; 39 in 1st practical year	history-taking and diagnosing
5	34	document-based (video vignettes)	Teacher Education: Physics & Biology	pre-service teachers Gender: 21f/12 m/1 Age: 24.06 (4.25) Semester: 4.79 (2.04)	diagnosing secondary school students' scientific reasoning skills
6	101	document-based (written material)	Medical Education	pre-service medical students Gender: 62f/39 m/1na Age: 29.68 (10.90) Semester: 8,39 (2.86)	diagnosing collaboratively with simulated radiologist
7	66	interaction-based (one-on-one interviews)	Teacher Education: Mathematics	pre-service teachers Gender: 38f/26 m/2na Age: 23.83 (5.74) Semester: 4.71 (2.49)	diagnosing 6th graders' understanding of decimal fractions

Note: Age and Semester *M*(*SD*) are provided.

choice from short and long lists). To ensure consistency in measuring participants' performance for this analysis, we focused exclusively on the accuracy of their diagnoses. The diagnoses were assessed as either *correct* or *incorrect* based on agreement with the expert solution for each simulated scenario, regardless of how they were entered or coded. The accuracy was then averaged across multiple scenarios that were diagnosed within the simulation-based learning environment.

Involvement. As no single scale existed that fit our needs, we constructed a scale with four items measuring perceived involvement based on prior scales focusing on cognitive involvement (Seidel et al., 2010; Vorderer et al., 2004), which was implemented in all studies. The scale uses a five-point Likert scale format (strongly agree; agree; undecided; disagree; strongly disagree). An example item is "While I was diagnosing in the simulated learning environment, I dedicated myself completely to the situation." The scale showed an internal consistency between $\alpha = 0.47$ and $\alpha = 0.79$ for the included studies.

Authenticity. As with involvement, we adapted items from prior authenticity scales (Schubert et al., 2001; Seidel et al., 2010). The five-point Likert scale (strongly agree; agree; undecided; disagree; strongly disagree) consists of three items. An example item is "The learning environment seemed to be just like a real professional situation." The scale showed an internal consistency between $\alpha = 0.74$ and $\alpha = 0.91$. For the included studies.

Extraneous cognitive load. We measured perceived extraneous cognitive load with a three-item scale (Eysink et al., 2009; Opfermann et al., 2010), using a five-point Likert scale format (strongly agree; agree; undecided; disagree; strongly disagree). An example item is "How easy or difficult is it for you to work in the learning environment." The scale showed an internal consistency between $\alpha = 0.48$ and $\alpha = 0.87$ for the included studies.

Full versions of all scales and scale statistics for all studies are provided in the open science repository (https://osf.io/ckeby/?view_only=e5a22ffb2e4a41b681c9c224036751d5).

3.3. Statistical analyses

Measurement invariance implies that using the same questionnaire with different groups measures the same construct in the same way (e.g., Millsap, 2011) and is therefore an essential prerequisite to answering our research questions. As running all possible sets of comparisons across the seven studies (120 possible sets) would result in an increased alpha error due to multiple testing, we applied an algorithm to limit the number of invariance tests necessary. Euclidean distances between the vectors of loadings estimated for each study individually were used to determine the distance in loadings. This allowed us to begin the invariance testing with the closest measures and then extend the testing incrementally based on this information (akin to the approach used for hierarchical cluster analysis; Bridges Jr., 1966). To investigate the research questions, it would be sufficient to establish metric (weak) invariance. Metric invariance ensures equivalence in measurement units by constraining factor loadings to be equal across groups. This constraint enables comparisons of factor covariances or unstandardized regression coefficients (Millsap, 2011). Invariance was assessed by evaluating the model fit, through chi-squared difference tests between the unconstrained (configural) and the metric model. Given the explorative nature

of our invariance analyses, a significance level of 1% was used for hypothesis testing.

To examine our research questions, we established structural equation models to analyze the correlations between the constructs. The latent correlations were then aggregated using the Fisher r-to-z transformed correlation coefficient, which served as the outcome measure. A random-effects model was fitted to the data, and the restricted maximum-likelihood estimator was used to estimate the amount of heterogeneity (τ^2) (Borenstein et al., 2009; Viechtbauer, 2010). Additionally, the *Q*-test for heterogeneity and the I^2 statistic were calculated and reported. The significance level for these analyses was set to 5%.

For correlations for which significant heterogeneity was detected, we tested whether the variation in study effects could be explained by the study domain or simulation type using categorical moderator analyses (Research Question 2). The significance level for these analyses was set to 5%. All analyses were conducted in R 4.0.3. The respective syntax can be found in the open science repository (https://osf.io/ckeby/?view_only=e5a22ffb2e4a41b681c9c224036751d5).

4. Results

4.1. Preliminary analysis: measurement invariance

Fig. 1 depicts the sets of studies for which we conducted analyses of invariance. Studies that are most similar in terms of the factor loadings for the respective constructs are clustered together. The height of the dendrograms indicates the similarity (distance) in studies. Less similar studies are successively grouped into larger clusters. The *p*-values indicate the likelihood of these studies being invariant with regard to their operationalization of the constructs. As can be seen in Fig. 1, we found metric measurement invariance across all studies for perceived authenticity and involvement. No acceptable measurement model could be fitted for involvement in Study 3, and thus the distance between this and other studies could not be determined. Accordingly, this study was excluded from further analyses.

Finally, perceived extraneous cognitive load did not show invariance across all studies, while metric invariance was found across all studies except Study 2. Accordingly, Study 2 was excluded from further analyses.

4.2. Learners' experience measures and diagnostic accuracy (RQ1)

The Fisher r-to-z transformed correlation coefficients for the correlation between perceived authenticity and perceived involvement ranged from 0.23 to 0.77, with all estimates being positive (100%). The estimated average Fisher r-to-z transformed correlation coefficient based on the random effects model was $\rho = 0.52$ (95% *CI:*.36 to 0.67, $p < 0.001$). According to the *Q*-test, the true correlations appear to be heterogeneous ($Q\ (5) = 11.69$, $p = 0.039$, $\tau^2 = 0.022$, $I^2 = 58.38\%$).

The observed Fisher r-to-z transformed correlation coefficients for the correlation between perceived authenticity and perceived extraneous cognitive load ranged from −0.80 to 0.11, with the majority of estimates being negative (80%). The estimated average Fisher r-to-z transformed correlation coefficient based on the random effects model was $\rho = -0.22$ (95% *CI:* −0.54 to 0.09, $p = 0.162$). According to the *Q*-test, the true outcomes appear to be heterogeneous ($Q\ (4) = 32.23$, $p < 0.001$, $\tau^2 = 0.112$, $I^2 = 88.93\%$).

The observed Fisher r-to-z transformed correlation coefficients for the correlation between perceived involvement and perceived extraneous cognitive load ranged from −0.72 to −0.01, with all the estimates being negative (100%). The estimated average Fisher r-to-z transformed correlation coefficient based on the random-effects model was $\rho = -0.40$ (95% *CI*: −0.63 to −0.16, $p = 0.001$). According to the *Q*-test, the true outcomes appear to be heterogeneous ($Q\ (4) = 19.54$, $p < 0.001$, $\tau^2 = 0.058$, $I^2 = 80.58\%$).

The observed Fisher r-to-z transformed correlation coefficients for the correlation between perceived authenticity and diagnostic accuracy ranged from −0.02 to 0.25, with the majority of estimates being positive (83%). The estimated average Fisher r-to-z transformed correlation coefficient based on the random effects model was $\rho = 0.13$ (95% *CI:*.03 to 0.23, $p = 0.010$). According to the *Q*-test, there was no significant amount of heterogeneity in the true outcomes ($Q\ (5) = 3.66$, $p = 0.599$, $\tau^2 = 0.000$, $I^2 = 0.00\%$).

The observed Fisher r-to-z transformed correlation coefficients for the correlation between perceived involvement and diagnostic accuracy ranged from −0.13 to 0.29, with the majority of estimates being positive (67%). The estimated average Fisher r-to-z transformed correlation coefficient based on the random effects model was $\rho = 0.12$ (95% CI: −0.01 to 0.25, $p = 0.067$). According to the *Q*-test, there was no significant amount of heterogeneity in the true outcomes ($Q\ (5) = 7.86$, $p = 0.164$, $\tau^2 = 0.010$, $I^2 = 37.82\%$).

Finally, the observed Fisher r-to-z transformed correlation coefficients for the correlation between perceived extraneous cognitive load and diagnostic accuracy ranged from −0.34 to −0.15, with all of the estimates being negative (100%). The estimated average Fisher r-to-z transformed correlation coefficient based on the random-effects model was $\rho = -0.22$ (95% *CI:* −0.32 to −0.12, $p < 0.001$). According to the *Q*-test, there was no significant amount of heterogeneity in the true outcomes ($Q\ (4) = 2.21$, $p = 0.697$, $\tau^2 = 0.000$, $I^2 = 0.00\%$). All aggregated correlations are summarized in Table 2.

4.3. Domains and simulation types as moderators (RQ2)

After calculating mean correlations between learners' experience measures and diagnostic accuracy, we tested whether the factors *domain* and/or *simulation type* could explain the heterogeneity in the correlations. Table 3 presents all aggregated correlations by domain and simulation type.

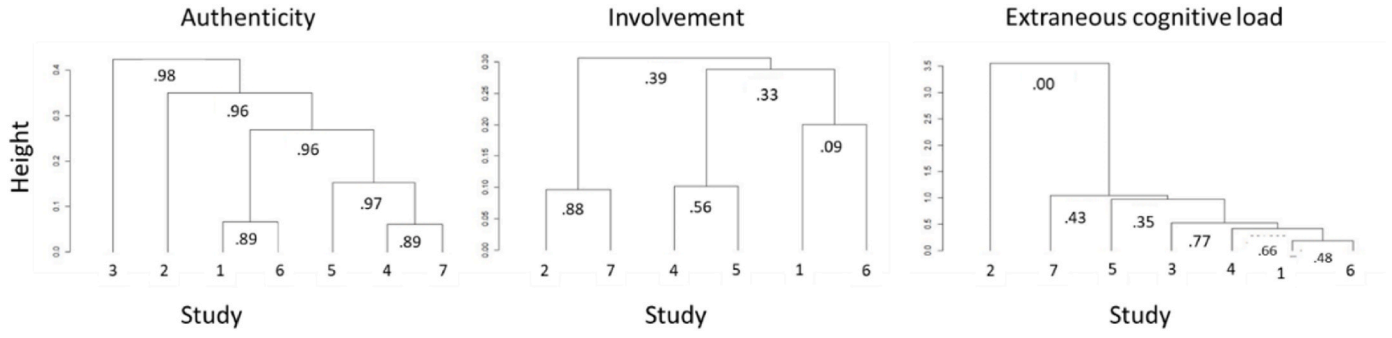

Fig. 1. *Dendrogram visualizing distances in loadings between all studies for authenticity, involvement, and extraneous cognitive load*
Note: Numbers represent p-values for Chi-Squared difference tests testing metric invariance across studies. $p > 0.010$ indicates non-significant results. Note that no measurement model for involvement could be fitted for Study 3.

Table 2
Aggregated correlations between learners' experience measures and diagnostic accuracy.

	Perceived authenticity	Perceived involvement	Perceived extraneous cognitive load
Perceived authenticity			
Perceived involvement	0.52 [0.36; 0.67]; $p < 0.001$		
Perceived extraneous cognitive load	−0.22 [-0.54; 0.09]; $p = 0.162$	−0.40 [-0.63; −0.16]; p =0 .001	
Diagnostic accuracy	0.13 [0.03; 0.23]; $p = 0.010$	0.12 [-0.01; 0.25]; $p = 0.067$	−0.22 [-0.32; −0.12]; p <0 .001

Note: numbers represent Fischer's *z* transformation of correlations, 95% confidence intervals, and *p*-values.

Table 3
Mean correlations between learners' experience measures and diagnostic accuracy for medical and teacher education and document vs. interaction-based simulations.

	Medical education	Teacher education	*p*
Perceived authenticity & Perceived involvement	0.58 [0.46; 0.77]; $p < 0.001$	0.55 [0.45; 0.63]; $p <. 001$	0.554
Perceived authenticity & Perceived extraneous cognitive load	−0.17 [-0.46; 0.11]; $p = 0.201$	−0.13 [-0.35; 0.08]; $p = 0.187$	0.909
Perceived involvement & Perceived extraneous cognitive load	−0.31 [-0.46; −0.13]; $p < 0.001$	−0.09 [-0.23; 0.05]; $p = 0.171$	0.828
Perceived authenticity & Diagnostic accuracy	0.11 [-0.16; 0.38]; $p = 0.292$	0.17 [-0.07; 0.41]; $p = 0.149$	0.907
Perceived involvement & Diagnostic accuracy	0.25 [0.11; 0.40]; $p < 0.001$	0.08 [-0.09; 0.24]; $p = 0.261$	0.014
Perceived extraneous cognitive load & Diagnostic accuracy	−0.20 [-0.38; −0.03]; $p = 0.028$	−0.17 [-0.33; −0.01]; $p = 0.046$	0.661
	Document-based	Interaction-based	*p*
Perceived authenticity & Perceived involvement	0.51 [0.46; 0.67]; $p < 0.001$	0.57 [0.41; 0.81]; $p < 0.001$	0.418
Perceived authenticity & Perceived extraneous cognitive load	−0.04 [-0.20; 0.12]; $p = 0.348$	−0.31 [-0.47; −0.15]; $p < 0.001$	0.004
Perceived involvement & Perceived extraneous cognitive load	−0.14 [-0.29; 0.02]; $p = 0.086$	−0.28 [-0.53; −0.02]; $p = 0.021$	0.562
Perceived authenticity & Diagnostic accuracy	0.19 [0.04; 0.35]; $p = 0.021$	0.05 [-0.09; 0.20]; $p = 0.602$	0.190
Perceived involvement & Diagnostic accuracy	0.15 [-0.01; 0.31]; $p = 0.072$	0.18 [0.07; 0.28]; $p = 0.090$	0.479
Perceived extraneous cognitive load & Diagnostic accuracy	−0.15 [-0.28; −0.03]; $p = 0.024$	−0.21 [−0.32; −0.01] $p = 0.004$	0.154

Note: numbers represent Fischer's *z* transformation of correlations, 95% confidence intervals, and p-values; *p* represents the p-values for the moderation effect.

Only the correlation between perceived involvement and diagnostic accuracy was moderated by the domain ($p = 0.014$), with stronger positive Fisher r-to-z transformed correlations for medical education ($\rho = 0.25$) than for teacher education ($\rho = 0.08$). Only the correlation between perceived authenticity and perceived extraneous cognitive load was moderated by the simulation type ($p = 0.004$), with stronger negative Fisher r-to-z transformed correlations for interaction-based simulations ($\rho = -0.31$) than for document based simulations ($\rho = -0.04$).

5. Discussion

The findings of this cross-study comparison show that learners' experience ratings are associated with objective performance measures (i.e. diagnostic accuracy). It is important to note that the assumption of measurement invariance for constructs related to learners' experience cannot be assumed without empirical validation. Failure to establish measurement invariance can significantly reduce the generalizability of findings (e.g., Vandenberg & Lance, 2000) and hinder meaningful comparisons across different studies. This study provides supportive evidence to such measurement invariance of some of the learners' experience scales, therefore, cross-domain comparisons were possible.

The results of this study suggest that the scales utilized to measure perceived authenticity and involvement have similar structures across domains and that the constructs can be measured equivalently in both medical and teacher education. In contrast, the measures of perceived extraneous cognitive load seem to be more specific to the simulation and situation. In other words, the availability of relevant information and the effort needed to deal with the features of the learning environment (Eysink et al., 2009) can be perceived differently in different contexts (e.g., different domains). Although metric invariance was reached, the clustering approach identified no patterns to explain variation based on the tested features.

The analyses of the relation between the constructs showed that perceived authenticity is closely related to perceived involvement across domains and types of simulations. As hypothesized, both of these measures are positively related to performance, as measured by diagnostic accuracy. In turn, perceived extraneous cognitive load is negatively related to other experience measures and diagnostic accuracy. Although these findings are correlational in nature, they are in line with previous research findings (e.g., Darling-Aduana, 2021; Seidel et al., 2010) which suggest that there might be a causal relation between perceived authenticity and involvement, which, in turn facilitates performance. This causal relationship needs to be tested in further primary studies. Our findings are also consistent with the discussion in Codreanu et al. (2020), emphasizing the essential role of authenticity and cognitive demand in designing simulation-based learning environments. According to our findings, perceived extraneous cognitive load (e.g., additional effort needed to deal with the learning environment) seems to be related to lower perceived authenticity and involvement as well as lower diagnostic accuracy. We can assume that lower perceived authenticity contributes to an increase in perceived extraneous cognitive load, as participants might think they are not working on core practical tasks. In contrast, higher perceived authenticity can relate to perceiving cognitive demand as an intrinsic cognitive load (e.g., see Sweller et al., 2019). However, this assumption needs to be validated by further evidence from primary research. If we consider authenticity as a design feature that can be manipulated (Chernikova et al., 2023), this finding supports the idea of balancing authenticity and cognitive demand (e.g., making some essential aspects of simulation authentic, while keeping other aspects simple to reduce possible distraction) in designing learning environments.

Our analysis also contributes to the argument by Chernikova et al. (2020) that the type of simulation (i.e., document-based vs. interaction-based simulation) might better explain the variance in the effects of simulations on performance and learning outcomes than domains alone. In this study, the negative relationship between perceived authenticity and perceived extraneous cognitive load was stronger for interaction-based simulations than for document-based simulations. One of the possible explanations for this phenomenon is that interaction-based simulations are perceived as more cognitively demanding, and the additional effort required to identify relevant information in interactions with simulated colleagues or non-professional interaction partners seems overwhelming. Moreover, the remaining unexplained variance indicates the presence of other possible moderators, such as professional knowledge, requiring further investigation.

There are a few limitations to mention. First, this study is exemplary, as it focuses only on one complex skill relevant to higher education (i.e., diagnostic skill) and three specific scales measuring perceived authenticity, perceived involvement, and perceived extraneous cognitive load. Accordingly, the findings cannot necessarily be generalized to other skills or scales measuring the same or similar constructs. Furthermore, the results can not be generalized to all possible types of simulations, for example, highly immersive simulations, simulations using embodied learning etc., as those learning environments might offer different levels of learners' experience. Based on our results, the scales we used demonstrated the ability to assess the intended constructs consistently across most contexts and domains. Building on this observation, we emphasize the importance of conducting invariance testing for the other scales used in this research. Second, rather methodological limitation, our results are based on relatively small sample sizes and a limited number of studies, which may be considered insufficient for a comprehensive meta-analytical approach. However, for the specific objective of this study, it was necessary to include a set of studies that investigated a similar phenomenon across different domains and contexts while utilizing identical measures. Given these extravagant constraints, our study achieved an acceptable sample size. Still, further evidence is needed based on primary research to better understand the impact of learners' experience on their development of complex skills. Third, this study did not examine possible changes in learners' experience (e.g., due to instructional support). Thus, further research and aggregation of intervention data are needed to better understand how learners experience changes within simulation-based learning environments and whether it has effects on their performance and learning outcomes. The use of process-related measures could lead to further insights, improving our understanding of the variation in experience measures and the effectiveness of simulations in developing diagnostic skills.

6. Conclusion

In conclusion, this study supports the importance of learners' experience ratings for assessing the effectiveness of simulations across different domains. Further, the study underscores the importance of conducting the measurement invariance analysis in cross-domain studies, ensuring the generalizability of the positive effects of instructional design that promote learners' involvement and authentic experiences. In addition, our findings offer supportive evidence for the theoretical models that highlight the relationship between learners' experience of simulated learning environments and learning or performance outcomes in simulation-based learning (Codreanu et al., 2020; Gegenfurtner et al., 2014; Seidel et al., 2010). Simulated learning environments hold significant potential for training complex skills in situations that are otherwise rare or dangerous (e.g., Morélot et al., 2021, Ziv et al., 2003). However, to fully harness this potential, it is crucial to appropriately combine authentic and immersive situations with adequately challenging learning tasks.

CRediT authorship contribution statement

Olga Chernikova: Writing – review & editing, Writing – original draft, Project administration, Data curation, Conceptualization. **Matthias Stadler:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Daniel Sommerhoff:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **Christian Schons:** Writing – original draft, Investigation, Data curation. **Nicole Heitzmann:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **Doris Holzberger:** Writing – review & editing, Writing – original draft, Conceptualization. **Tina Seidel:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Data curation, Conceptualization. **Constanze Richters:** Writing – original draft, Investigation. **Amadeus J. Pickal:** Writing – original draft, Investigation. **Christof Wecker:** Supervision, Funding acquisition, Conceptualization. **Michael Nickl:** Writing – original draft, Investigation, Conceptualization. **Elias Codreanu:** Writing – original draft, Investigation, Conceptualization. **Stefan Ufer:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization. **Stephanie Kron:** Writing – original draft, Investigation. **Caroline Corves:** Writing – original draft, Investigation. **Birgit J. Neuhaus:** Writing – original draft, Supervision, Funding acquisition, Data curation. **Martin R. Fischer:** Writing – original draft, Supervision, Funding acquisition, Conceptualization. **Frank Fischer:** Writing – review & editing, Writing – original draft, Supervision, Resources, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements and funding

The research for this article was funded by the German Research Association (Deutsche Forschungsgemeinschaft, DFG): DFG FOR2385; FI792/12.

References

Al-Kadi, A. S., & Donnon, T. (2013). Using simulation to improve the cognitive and psychomotor skills of novice students in advanced laparoscopic surgery: A meta-analysis. *Medical Teacher, 35*(Suppl 1), 47–55. https://doi.org/10.3109/0142159X.2013.765549

Artelt, C., & Rausch, T. (2014). Accuracy of teacher judgments. In S. Krolak-Schwerdt, S. Glock, & M. Böhmer (Eds.), *Teachers' professional development: Assessment, training, and learning. The future of education research* (Vol. 3, pp. 27–43). Sense Publishers.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies. *Zeitschrift für Psychologie, 223*, 3–13. https://doi.org/10.1027/2151-2604/a000194

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Bridges, C. C. (1966). Hierarchical cluster analysis. *Psychological Reports, 18*(3), 851–854. https://doi.org/10.2466/pr0.1966.18.3.851

Brom, C., Děchtěrenko, F., Frollová, N., Stárková, T., Bromová, E., & D'Mello, S. K. (2017). Enjoyment or involvement? Affective-Motivational mediation during learning from a complex computerized simulation. *Computers & Education, 144*, 236–254. https://doi.org/10.1016/j.compedu.2017.07.001

Bygstad, B., Øvrelid, E., Ludvigsen, S., & Dæhlen, M. (2022). From dual digitalization to digital learning space: Exploring the digital transformation of higher education. *Computers & Education, 182*. https://doi.org/10.1016/j.compedu.2022.104463. e-104463.

Charlin, B., Tardif, J., & Boshuizen, H. P. A. (2000). Script and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research. *Academic Medicine: Journal of the Association of American Medical Colleges, 75*(2), 182–190.

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research, 90*(4), 499–541. https://doi.org/10.3102/0034654320933544.

Chernikova, O., Holzberger, D., Heitzmann, N., Stadler, M., Seidel, T., & Fischer, F. (2023). Where salience goes beyond authenticity: A meta-analysis on simulation-based learning in higher education. *Zeitschrift für Pädagogische Psychologie, 38*(1–2), 15–25. https://doi.org/10.1024/1010-0652/a000357.

Codreanu, E., Sommerhoff, D., Huber, S., Ufer, S., & Seidel, T. (2020). Between authenticity and cognitive demand: Finding a balance in designing a video-based simulation in the context of mathematics teacher education. *Teaching and Teacher Education, 95*, Article 103146. https://doi.org/10.1016/j.tate.2020.103146

Cook, D. A. (2014). How much evidence does it take? A cumulative meta-analysis of outcomes of simulation-based education. *Medical Education, 48*(8), 750–760. https://doi.org/10.1111/medu.12473

Dankbaar, M. E. W., Alsma, J., Jansen, E. E. H., Merrienboer, J. J. G., van Saase, J. L. C. M., & Schuit, S. C. E. (2016). An experimental study on the effects of a simulation game on students' clinical cognitive skills and motivation. *Advances in Health Sciences Education, 21*(1), 505–521. https://doi.org/10.1007/s10459-015-9641-x

Darling-Aduana, J. (2021). Authenticity, engagement, and performance in online high school courses: Insights from micro-interactional data. *Computers & Education, 167*, Article 104175. https://doi.org/10.1016/j.compedu.2021.104175

Eysink, T. H. S., de Jong, T., Berthold, K., Kolloffel, B., Opfermann, M., & Wouters, P. (2009). Learner performance in multimedia learning arrangements: An analysis across instructional approaches. *American Educational Research Journal, 46*(4), 1107–1149. https://doi.org/10.3102/0002831209340235

Fink, M. C., Radkowitsch, A., Bauer, E., Sailer, M., Kiesewetter, J., Schmidmaier, R., Siebeck, M., Fischer, F., & Fischer, M. (2021). Simulation research and design: A dual-level framework for multi-project research programs. *Educational Technology Research & Development, 69*, 809–841. https://doi.org/10.1007/s11423-020-09876-0

Fischer, F., Chernikova, O., & Opitz, A. (2022). Learning to Diagnose with Simulations: Introduction. In F. Fischer, & A. Opitz (Eds.), *Learning to Diagnose with Simulations*. Cham: Springer. https://doi.org/10.1007/978-3-030-89147-3_1.

Frasson, C., & Blanchard, E. G. (2012). Simulation-based learning. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 3076–3080). Springer. https://doi.org/10.1007/978-1-4419-1428-6_129.

Gartmeier, M., Bauer, J., Fischer, M. R., Hoppe-Seyler, T., Karsten, G., Kiessling, C., Möller, G. E., Wiesbeck, A., & Prenzel, M. (2015). Fostering professional communication skills of future physicians and teachers: Effects of e-learning with video cases and role-play. *Instructional Science, 43*(4), 443–462. https://doi.org/10.1007/s11251-014-9341-6

Gegenfurtner, A., Quesada-Pallarès, C., & Knogler, M. (2014). Digital simulation-based training: A meta-analysis. *British Journal of Educational Technology, 45*(6), 1097–1114. https://doi.org/10.1111/bjet.12188

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. W. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 111*(9), 2055–2100.

Hege, I., Kononowicz, A. A., Kiesewetter, J., & Foster-Johnson, L. (2018). Uncovering the relation between clinical reasoning and diagnostic accuracy–An analysis of learner's clinical reasoning processes in virtual patients. *PLoS One, 13*(10), Article e0204900. https://doi.org/10.1371/journal.pone.0204900

Heitzmann, N., Opitz, A., Stadler, M., Sommerhoff, D., Fink, M. C., Obersteiner, A., Schmidmaier, R., Neuhaus, B. J., Ufer, S., Seidel, T., Fischer, M. R., & Fischer, F. (2021). Cross-disciplinary research on learning and instruction – coming to terms. *Frontiers in Psychology, 11*, Article e562658. https://doi.org/10.3389/fpsyg.2021.562658

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M., Ufer, S., Schmidmaier, R., Neuhaus, B., Siebeck, M., Stürmer, K., Obersteiner, A., Reiss, K., Girwidz, R., & Fischer, F. (2019). Facilitating diagnostic competences in simulations: A conceptual framework and a research agenda for medical and teacher education. *Frontline Learning Research, 7*(4), 1–24. https://doi.org/10.14786/flr.v7i4.384

Helmke, A., Schrader, F.-W., & Helmke, T. (2012). EMU: Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung. Unterrichtsdiagnostik – Ein Weg, um Unterrichten sichtbar zu machen. *Schulverwaltung Bayern, 35*(6), 180–183. https://journals.lww.com/academicmedicine/Fulltext/2000/02000/Scripts_and_Medical_Diagnostic_Knowledge__Theory.20.aspx.

Ijsselsteijn, W. A., & Riva, G. (2003). Being there: The experience of presence in mediated environments. In G. Riva, F. Davide, & W. A. IJsselsteijn (Eds.), *Being there: Concepts, effects and measurements of user presence in synthetic environments* (Vol. 5, pp. 3–16). IOS Press.

Lesā, R., Daniel, B., & Harland, T. (2021). Learning with simulation: The experience of nursing students. *Clinical Simulation in Nursing, 56*, 57–65. https://doi.org/10.1016/j.ecns.2021.02.009

Millsap, R. (2011). *Statistical approaches to measurement invariance*. Routledge/Taylor & Francis Group.

Morélot, S., Garrigou, A., Dedieu, J., & N'Kaoua, B. (2021). Virtual reality for fire safety training: Influence of immersion and sense of presence on conceptual and procedural acquisition. *Computers & Education, 166*. https://doi.org/10.1016/j.compedu.2021.104145. e-104145.

Opfermann, M., Scheiter, K., & Gerjets, P. (2010). Hypermedialeren: De invloed van instructieontwerp, leerlingkenmerken en ondersteuning [Hypermedia learning: The impact of instructional design, learner characteristics, and instructional support]. *Pedagogische Studies, 87*, 9–26.

Palmer, M. T. (1995). Interpersonal communication and virtual reality: Mediating interpersonal relationships. In F. Biocca, & M. R. Levy (Eds.), *Communication in the age of virtual reality* (pp. 277–302). Lawrence Erlbaum.

Reeves, S. M., Crippen, K. J., & McCray, E. D. (2021). The varied experience of undergraduate students learning chemistry in virtual reality laboratories. *Computers & Education, 175*, e–104320. https://doi.org/10.1016/j.compedu.2021.104320

Robitzsch, A., & Lüdtke, O. (2020). A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psychological Test and Assessment Modeling, 62*(2), 233–279. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-2/03_Robitzsch.pdf.

Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Presence: Teleoperators and Virtual Environments, 10*(3), 266–281. https://doi.org/10.1162/105474601300343603

Seidel, T., Blomberg, G., & Stürmer, K. (2010). OBSERVE - Validierung eines videobasierten Instruments zur Erfassung der professionellen Wahrnehmung von Unterricht. In E. Klieme, D. Leutner, & M. Kenk (Eds.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. 56. Beiheft der Zeitschrift für Pädagogik, Weinheim u.a* (pp. 296–306). Beltz.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. *Zeitschrift für Padagogische Psychologie, 19*(1), 85–95.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(1), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Theelen, H., Van den Beemt, A., & den Brok, P. (2019). Classroom simulations in teacher education to support preservice teachers' interpersonal competence: A systematic literature review. *Computers & Education, 129*(1), 14–26. https://doi.org/10.1016/j.compedu.2018.10.015

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70. https://doi.org/10.1177/109442810031002

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Vorderer, P., Wirth, W., Ribeiro Gouveia, F., Biocca, F., Saari, T., Jäncke, L., Böcking, S., Schramm, H., Gysbers, A., Hartmann, T., Klimmt, C., Laarni, J., Ravaja, N., Sacau, A., Baumgartner, T., & Jäncke, P. (2004). MEC Spatial Presence Questionnaire (MEC-SPQ) short documentation and instructions for application. http://academic.csuohio.edu/kneuendorf/frames/MECFull.pdf.

Wirth, W., Hartmann, T., Böcking, S., Vorderer, P., Klimmt, C., Schramm, H., Saari, T., Laarni, J., Ravaja, N., Gouveia, F. R., Biocca, F., Sacau, A., Jäncke, L., Baumgartner, T., & Jäncke, P. (2007). A process model of the formation of spatial presence experiences. *Media Psychology, 9*(3), 493–525. https://doi.org/10.1080/15213260701283079

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments, 7*, 225–240. https://doi.org/10.1162/105474698565686

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences, 45*, e1. https://doi.org/10.1017/S0140525X20001685

Ziv, A., Wolpe, P. R., Small, S. D., & Glick, S. (2003). Simulation-based medical education: An ethical imperative. *Academic Medicine, 78*(8), 783–788.