Editorial

Validly Authentic

Some Recommendations to Researchers Using Simulations in Psychological Assessment

Matthias Stadler[1], Dragos Iliescu[2], and Samuel Greiff[3]

[1] Department of Psychology, Ludwig-Maximilians-University of Munich, Germany
[2] Faculty of Psychology and Educational Sciences, University of Bucharest, Romania
[3] Department of Behavioural and Cognitive Sciences, University of Luxembourg, Luxembourg

Technology has become an indispensable tool for educational and psychological assessment in today's world (Goldhammer et al., 2019). Digital assessments facilitate the creation of new types of stimuli and response formats that were out of reach for assessments using paper; for instance, interactive simulations may include multimedia elements, as well as virtual or augmented realities (Cipresso et al., 2018; de-Juan-Ripoll et al., 2018). These types of assessments also allow for widening the construct coverage in an assessment, for instance, through stimulating and making visible specific problem-solving strategies that represent new forms of problem-solving (Han et al., 2019). Correspondingly, simulation-based assessment tools that authentically represent certain aspects of reality but remain subject to the control of users (researchers and practitioners) have long been proposed to bridge the gap between laboratory and field assessment (Brehmer & Dörner, 1993).

Indeed, simulations have proven to be potent tools in assessment and training across many different domains (Chernikova et al., 2020), with a primary contribution to the assessment's realism and ecological validity. Unlike more traditional assessment instruments, simulations often emulate real-world problem-solving scenarios attempting to capture highly complex abilities such as medical reasoning (Schuwirth & van der Vleuten, 2003) or diagnostic competencies (Chernikova et al., 2020). Striving for authentic simulations harbors the danger of losing focus of the intended construct by introducing additional challenges that require different competencies to solve the problems posed by the simulations, such as handling the interface or, more generally, competencies that are not construct-relevant but rather relate to, for instance, human-computer interaction. In other words, the authenticity of a simulation may add systematic variance to the measure that is not necessarily due to the intended construct.

In this Editorial, we encourage researchers to develop complex simulations to assess novel and exciting constructs without losing track of their targeted constructs when designing authentic simulations (Greiff, 2017). We will also focus on the pervasive difficulty of adequate consideration of participants' previous knowledge in measures of ability (such as general or domain-specific problem solving).

The Interplay Between Knowledge and Ability in Simulations

To illustrate the interplay between knowledge and ability in simulations, take this example. A highly domain-general simulation might ask a participant to determine the relation between a set of fictional drugs and some fictional diseases by interacting with the simulation (see, e.g., Greiff et al., 2015). In this case, medical knowledge does not help succeed in the task, as all features are fictional. However, critics could affirm that such a simulation is not authentic enough – and it would be effortlessly possible to shift this simulation toward an authentic medical scenario in which the names of drugs and diseases and their relations present realistic cases (see, e.g., Drews et al., 2006). In this case, the construct assessed would shift from domain-general reasoning to domain-specific (medical) reasoning.

Suppose the task becomes maximally realistic and requires expert knowledge to be solved. In that case, it will also assess specific domain knowledge rather than a latent problem-solving ability (e.g., it will implicitly contain a measure of medical knowledge on a specific topic). This interaction between knowledge, ability, and problem-solving performance has been described in the Elshout-Raaheim-Hypothesis

(Elshout, 1987; Raaheim, 1988; see also Leutner, 2002; Weise et al., 2020). The hypothesis describes a non-linear moderation of the relation between ability and performance by applicable knowledge. If a participant knows how to solve the problem in a simulation, the problem becomes a mere task and does not require any ability.

On the other hand, if the participant does not have sufficient knowledge to interact with the problem, they have no opportunity to use their ability (i.e., mere guessing). In the example introduced above, the same simulation could be trivial to a medical expert but virtually unsolvable to a novice. Therefore, a simulation intended to assess any latent ability should require just enough domain knowledge so the target population will interact systematically with the simulation without knowing the correct solution at the simulation's onset.

What Is Measured?

Making sure the simulation assesses the intended construct requires careful item construction. The principles we want to stress in this Editorial are not unique to simulation-based assessment (see Ziegler, 2014) but become particularly apparent in these cases where validity and authenticity need to be balanced. In particular, we want to focus on four aspects of assessment development: (1) clear definitions of the targeted construct and (2) the targeted population as well as (3) theoretical considerations of adequate difficulty, and (4) required number of simulations.

First, there should be a clear definition of the construct the assessment intends to measure. This call for the definition of constructs may seem trivial, but many simulations prioritize authentic representation of real-life problem scenarios over psychometric quality and psychological theory (Codreanu et al., 2020; Schuwirth & van der Vleuten, 2003). Next to a theoretical description of the construct itself should entail a description of the construct's nomological network (i.e., the construct's relation to observables and different theoretical constructs; Cronabch & Meehl, 1955). For research on simulations submitted to *EJPA*, we expect a concise definition of the construct, as with all research submitted to *EJPA*, but will specifically focus on the trade-off between authenticity and validity.

Second, because of the non-linear interaction between domain knowledge, ability, and problem-solving performance described above, the target population needs to be clearly defined regarding their expected level of domain knowledge. The intended target population is, thus, closely linked to the intended use of the simulation (Ziegler, 2014). A simulation that provides a good assessment for novices may not apply to experts or vice versa. When submitting research to *EJPA*, please state the intended use of your simulation, the target population, and potential issues with using the simulation outside of the target population. It should also be evident that the samples used to construct the test must be derived from the targeted population.

Third, the concept of adequacy for the target population leads directly to the question of item difficulty. Ideally, researchers have a theoretical model of what features of their items influence difficulty (Gierl & Haladyna, 2012). This model should include domain-specific aspects in a simulation emulating real-world problem-solving scenarios such as domain knowledge and domain-general aspects (Schuwirth & van der Vleuten, 2003). These features of simulations are theoretically orthogonal as it is possible to make a problem more difficult by increasing domain-specific features independently of domain-general features. Applied to the example of the medical simulation introduced above, the difficulty could be increased by using rare or specific diseases (Norman et al., 2006) or by adding complexity by increasing the number of variables to consider or the dynamics of the simulation (e.g., Stadler et al., 2016). These features are generally independent of each other, and researchers should theoretically describe their contribution to task difficulty before constructing the simulations. A sufficiently elaborated theoretical model can be tested empirically using item-response approaches such as the linear logistic test model (Kubinger, 2009). Not every submission of simulation-based assessment to *EJPA* will need to provide empirical tests of the underlying item model. However, there should be theoretical considerations as to what features of the simulation contribute to item difficulty and some reasoning as to why the exact set of simulations was chosen beyond authentic content. Finally, testing a theoretical model of the impact had by the features of a simulation on item difficulty requires employing a whole set of systematically constructed simulations that assess the same ability, which leads to the final aspect that we want to emphasize: the required amount of different (and at least somewhat independent) simulations in an assessment. This is another trade-off situation in which validity can easily be overshadowed by convenience. Developing authentic simulations is often very laborious and costly for the test developer, and complex simulations may also take test-takers more time to complete. For these reasons, researchers are often limited in the number of simulations they can include in their assessment. It is essential to consider that each simulation represents a single item in a test providing just a single indicator of the latent ability of interest. Even if multiple behavior indicators are extracted from a simulation, these are highly interdependent. The dependency should be accounted for statistically by applying hierarchical models

(e.g., von Davier et al., 2017) or aggregating all indicators to a single performance score.

Using a small number of simulations (or even just a single simulation) to assess a latent construct may lead to unreliable results. Especially highly context-rich simulations are likely to have considerable measurement error. The contexts of the various simulations will trigger different expectations in participants resulting in different answering behaviors and differential item functioning (Holland & Wainer, 2012). Examples for this are plentiful in the early days of assessing complex problem solving using simulations (Stadler et al., 2015). These first simulations emulated very complex problems (such as running a whole town) consisting of hundreds of variables and requiring multiple days of testing (Doerner, 1980). Correspondingly, they employed only one unique simulation with questionable psychometric quality (Süß et al., 1993).

In deciding how many simulations to use for an assessment, we encourage researchers to apply the same criteria as any other psychometric measure. In line with the previous statements on simulation difficulty, this also includes an adequate distribution of difficulty across the assessment to allow for discrimination between groups of low, average, and high performers (Greiff et al., 2015). For every submission to *EJPA*, we will critically evaluate whether the authors provide convincing arguments for the length of their assessment instrument.

Conclusion

Simulations allow researchers to assess essential abilities under authentic conditions yielding the potential to increase, among other, construct coverage. Moreover, the complexity of some abilities (such as complex or collaborative problem solving) cannot be realized at all without simulations. In this Editorial, we want to encourage authors to continue exploring the use of simulations as assessment tools but also make them aware of potential pitfalls that come with them. Most importantly, researchers need to adhere to the standards of psychometric quality when using simulations in assessment. Suppose the simulation requires too much time or is too costly to employ sufficient different simulations for an acceptably reliable measurement, it may not be the adequate approach to measure the intended construct. However, if researchers use simulations appropriately, they represent potent tools that allow for deep insights into psychological processes that are not accessible to psychometric assessment. So please submit your research on simulations in assessment to *EJPA*, including adequate sections on all four issues raised in this Editorial.

References

Brehmer, B., & Dörner, D. (1993). Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Computers in Human Behavior, 9*(2–3), 171–184. https://doi.org/10.1016/0747-5632(93)90005-d

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

Cipresso, P., Pedroli, E., Serino, S., Semonella, M., Tuena, C., Colombo, D., Pallavicini, F., & Riva, G. (2018). Assessment of unilateral spatial neglect using a free mobile application for Italian clinicians. *Frontiers in Psychology, 9*, Article 2241. https://doi.org/10.3389/fpsyg.2018.02241

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

Cronabch, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

de-Juan-Ripoll, C., Soler-Domínguez, J. L., Guixeres, J., Contero, M., Álvarez Gutiérrez, N., & Alcañiz, M. (2018). Virtual reality as a new approach for risk taking assessment. *Frontiers in Psychology, 9*, Article 2532. https://doi.org/10.3389/fpsyg.2018.02532

Doerner, D. (1980). On the difficulties people have in dealing with complexity. *Simulation & Games, 11*(1), 87–106. https://doi.org/10.1177/104687818001100108

Drews, F. A., Syroid, N., Agutter, J., Strayer, D. L., & Westenskow, D. R. (2006). Drug delivery as control task: Improving performance in a common anesthetic task. *Human Factors, 48*(1), 85–94. https://doi.org/10.1518/001872006776412216

Elshout, J. J. (1987). Problem solving and education. In E. de Corte, H. Lodewijks, & R. Parmentier (Eds.), *Learning and instruction* (pp. 259–273). Pergamon.

Gierl, M. J., & Haladyna, T. M. (2012). *Automatic item generation*. Routledge. https://doi.org/10.4324/9780203803912

Goldhammer, F., Scherer, R., & Greiff, S. (2019). Editorial: Advancements in technology-based assessment: Emerging item formats, test designs, and data sources. *Frontiers in Psychology, 10*, Article 3047. https://doi.org/10.3389/fpsyg.2019.03047

Greiff, S. (2017). The field of psychological assessment: Where it stands and where it's going – A personal analysis of foci, gaps, and implications for *EJPA*. *European Journal of Psychological Assessment, 33*(1), 1–4. https://doi.org/10.1027/1015-5759/a000412

Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2015). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning, 21*(3), 356–382. https://doi.org/10.1080/13546783.2014.989263

Han, Z., He, Q., & von Davier, M. (2019). Predictive feature generation and selection using process data from PISA interactive problem-solving items: An application of random forests. *Frontiers in Psychology, 10*, Article 2461. https://doi.org/10.3389/fpsyg.2019.02461

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge. https://doi.org/10.4324/9780203357811

Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement, 69*(2), 232–244. https://doi.org/10.1177/0013164408322021

Leutner, D. (2002). The fuzzy relationship of intelligence and problem solving in computer simulations. *Computers in Human Behavior, 18*(6), 685–697. https://doi.org/10.1016/s0747-5632(02)00024-9

Norman, G., Bordage, G., Page, G., & Keane, D. (2006). How specific is case specificity? *Medical Education, 40*(7), 618–623. https://doi.org/10.1111/j.1365-2929.2006.02511.x

Raaheim, K. (1988). Intelligence and task novelty. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 4, pp. 73–97). Erlbaum.

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2003). The use of clinical simulations in assessment. *Medical Education, 37*(Suppl 1), 65–71. https://doi.org/10.1046/j.1365-2923.37.s1.8.x

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

Stadler, M., Niepel, C., & Greiff, S. (2016). Easily too difficult: Estimating item difficulty in computer simulated microworlds. *Computers in Human Behavior, 65*, 100–106. https://doi.org/10.1016/j.chb.2016.08.025

Süß, H.-M., Kersting, M., & Oberauer, K. (1993). Zur Vorhersage von Steuerungsleistungen an Computersimulierten Systemen durch Wissen und Intelligenz [On the predictability of control performance on computer-simulated systems by knowledge and intelligence]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 14*(3), 189–203.

von Davier, A. A., Zhu, M., & Kyllonen, P. C. (2017). *Innovative assessment of collaboration: Methodology of educational measurement and assessment*. Springer.

Weise, J. J., Greiff, S., & Sparfeldt, J. R. (2020). The moderating effect of prior knowledge on the relationship between intelligence and complex problem solving – testing the Elshout-Raaheim hypothesis. *Intelligence, 83*, Article 101502. https://doi.org/10.1016/j.intell.2020.101502

Ziegler, M. (2014). Stop and state your intentions. *European Journal of Psychological Assessment, 30*(4), 239–242. https://doi.org/10.1027/1015-5759/a000228

Published online December 6, 2021

Funding

This research was supported by a grant from the Deutsche Forschungsgemeinschaft DFG (COSIMA; DFG-Forschungsgruppe 2385).

Matthias Stadler
Ludwig-Maximilians-University of Munich
Leopoldstr. 13
80802 Munich
Germany
matthias.stadler@lmu.de

