A Commentary on Construct Validity when using Operational Virtual Learning Environment Data in Effectiveness Studies

A. Corinne Huggins-Manley
University of Florida

Carole R. Beal
University of Florida

Sidney K. D'Mello
University of Colorado Boulder

Walter L. Leite
University of Florida

Dyugu Dee Cetin-Berber
University of Florida

Dongho Kim
University of Florida

Danielle S. McNamara
Arizona State University

Corresponding author: A. Corinne Huggins-Manley; 1215 Norman Hall, Gainesville, FL 32611; (352) 273- 4342; amanley@coe.ufl.edu

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C160004 to the University of Florida. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

Virtual learning environments (VLE) are increasingly used at-scale in educational contexts to facilitate teaching and promote learning, and the data they produce can be used for educational research purposes. Meanwhile, the U. S. Department of Education's Office of Educational Technology has repeatedly emphasized the importance of using evidence to validate claims from VLE-based educational research. Although VLE data can provide some affordances for conducting educational research, we argue that many challenges can arise with respect to providing evidence for construct validity. The objective of this commentary is to encourage educational researchers using operational, at-scale VLE data to align their data and intended constructs to a theoretical framework of construct validity threats in order to develop a comprehensive set of actionable solutions. We use examples from our research project as a demonstration resource for performing such an alignment.

Introduction

Virtual learning environments (VLEs) are systems in which learners and teachers, or tutors, interact online for an educational purpose (Weller, 2007). VLEs are increasingly used at-scale in educational contexts to facilitate teaching and promote learning, and the data they produce can be used for educational research purposes (U. S. Department of Education's Office of Educational Technology, 2012, 2013). Operational data produced from at-scale VLE users adds to the "digital ocean," as termed by DiCerbo and Behrens (2014) and defined as "the vast amount of [educational] data that is available from interactions with digital tools" (pp. Preface). This type of data is quite different from "traditional" education data as it is often unstructured (National Forum on Educational Statistics, 2015), meaning it is numerous, fine-grained, on-going, scalable, and not derived from a pre-defined model. Cope and Kalantzis (2016) note that unstructured data from technology-mediated learning environments inherently involve data points that do not have an "immediate obvious meaning" (pp. 6). As aptly stated in a U. S. Department of Education (2013) report focused on improving evidence-based effectiveness research in a digital world, "Data do not interpret themselves" (pp. viii). However, educational effectiveness research studies that utilize VLE operational data must assign meaning and interpretations to such data in order to draw inferences about educational constructs. Such assignment of meaning to these unstructured data requires evidence of construct validity.

Construct validity is defined as the ability to make evidence supported inferences from sampled indicators to the constructs they are intended to represent (Shadish, Cook, & Campbell, 2002). As an example, a construct of interest in a VLE-based research project may be "student engagement" and example indicators may be click streams, eye tracking data (D'Mello, Dieterle, & Duckworth, 2017), verbal responses (McNamara et al., in press), or other software log files selected and manipulated via data mining techniques (Baker & Ocumpaugh, 2014). This example raises the following construct validity question: “What evidence do we have to infer an educational construct of “student engagement” from eye tracking patterns, clicks, verbal responses, and exploited log files?” As this example illustrates, although VLE-produced data can provide affordances for conducting educational research (e.g., Klasnia-Milicevic, Ivanovic, & Budimac, 2017; DiCerbo & Behrens, 2014; Susnea, 2018), we argue that many challenges can arise with respect to providing evidence for construct validity.

A core reason for this argument is that effectiveness research based on operational, at-scale VLE data often utilizes a reverse construct measurement process when compared to traditional educational research. In the latter, best practices entail first defining a construct of interest for a research project and then collecting data to represent that construct according to an *a priori* model (Shadish et al., 2002). But in research that utilizes operational VLE log files, the data has often been collected prior to the establishment of constructs for a particular research study and without any particular aim for collecting data in a manner that lends itself to educational constructs. This is because the primary goal of VLEs is to promote learning not to collect data for research studies. For quantifying constructs from the operational VLE data, the goal then becomes to extract and exploit the VLE data to construct variables, which can then be tested for the degree to which they adequately represent an educational construct (Baker & Yacef, 2009). One of several limitations to this reversed construct measurement process is that many VLEs used at scale are based on fairly simply design features that lack integration of innovations required for sophisticated construct extraction (Baker, 2016). But more importantly, this reversed measurement process contradicts the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, &

National Council on Measurement in Education, 2014), as well as seminal theoretical frameworks of study inference validity (e.g., Shadish et al., 2002).

In VLE-based educational research, the question of weak construct validity has been raised from a variety of angles, including (a) researchers proposing measurement methodologies (e.g., Baker, Ocumpaugh, & Andres, 2018; D'Mello, Kappa, & Gratch, 2017; Kai et al., 2018), (b) researchers discussing a host of methodological challenges in technology-based educational research (e.g., Daniel, 2015; Shute & Moore, 2017; U. S. Department of Education's Office of Educational Technology, 2013), and (c) researchers asking substantive questions about the effects of VLEs on various outcomes (e.g., Garcia-Alvareza, Novo-Cortib, & Varela-Candamioc, 2018; Idemudia & Negash, 2012). Nonetheless, we have not found any VLE-based educational research projects that report a thorough evaluation of their project in alignment with a theoretical framework of construct validity threats, which makes it difficult for readers to evaluate the evidentiary support for claims of educational effectiveness in studies utilizing operational VLE data. This is a critical issue that should take priority in educational research for at least two reasons: a) the proliferation of research based on operational VLE data, and b) the critical importance of using evidence to validate claims from VLE-based educational research (U. S. Department of Education's Office of Educational Technology, 2012, 2013).

Objective

The objective of this commentary is to encourage educational researchers using operational, at-scale VLE data to align their data and intended constructs to a theoretical framework of construct validity threats, a process which we argue allows researchers to develop actionable solutions to the complex and overwhelming challenge of construct validation. We use examples from our research project as a demonstration resource for performing such an alignment. Specifically, we align some experiences in our VLE-based research project to construct validity threats under the Shadish et al. (2002) framework. The larger project (Author) entails a series of research studies, all of which use logged data within an online algebra tutoring system for high school students in (State). Each school year, the logged data stems from hundreds of thousands of students and their teachers who interact with the VLE to view tutorial videos, engage with discussion boards, practice test items, use learning materials, and more. In (State), usage of the VLE is largely self-selected by teachers and students, and the VLE offers students open access such that new students can log in at any time. The curriculum and practice test items in the VLE were designed to have strong alignment to algebra curriculum standards. Notably, the system was designed for educational practice more so than research.

Examples

Table 1 displays examples of aligning our VLE-based research constructs to formal construct validity threats, with citations that illustrate how different features of our examples connect to the relevant literature. Readers should acknowledge that this type of table could be developed and consistently updated within their own research projects that use operational VLE data. The first column is a list of threats adapted from Shadish et al.'s (2002) framework for validity of causal inferences drawn from quantitative research studies. The second column provides a brief description of the presence of each of the threats. The third column provides *one* broad reason (of many possible reasons) that educational researchers using operational, at-scale VLE data should be considering, reflecting on, and evaluating the particular construct threat. This third column is a brief attempt to remind readers that while the threats in Shadish et al.'s (2002) framework were not developed with VLE-based research in mind, they remain relevant to such research. The fourth column provides *one* example from our project in which we have had concerns about the particular validity threat.

Table 1

Aligning Challenges to Construct Validity in Our VLE-based Research Project to Formal Validity Threats

Construct Validity Threat[1]	**Description of the *Presence* of the Threat[1]**	***One* Broad Reason for Considering this Threat in VLE-based Research**	***One* Example Threat from our VLE-based Research Project**
Mono-operation bias	A single indicator of a construct underrepresents the inferred construct, which is more complex than a single indicator.	Because data collection often occurs before the formal establishment of constructs[2], only one indicator of the construct may be available.	In order to conduct our project studies, we need estimates of the students' ongoing and changing algebra abilities throughout the interaction with the VLE. However, the non-standardized, self-selected administration of a few test items with simple scoring[3] provides only a single indicator per time point of a very complex construct, "student algebra ability."
Mono-method bias	The method becomes a part of the construct because only one method was used to measure it.	Nearly all the observed data comes from computer logs[2], a single data collection method.	"Student engagement" in our research studies is inferred through statistical techniques that make use of data stemming solely from computer-logged indicators (e.g., time logs, video selections).
Inadequate explication of construct	A construct defined in a vague or incomplete manner leads to invalid inferences about the construct.	Because data collection often occurs before the formal establishment of constructs[2], intended construct explications may not be aligned with the available data.	Early in our project, we measured "student usage" through the available computer-logged information residing under a student's login (e.g., student login frequency; video viewing; responding to practice test items). However, when interpreting and writing about our research, we were implicitly referring to the broader definition of the construct, that of the degree to which a student has been exposed to the materials in the VLE platform. This is problematic because, in our project, teacher surveys indicated that much of the student use of the VLE platform occurs in a classroom setting (e.g., teacher showing the videos during class), which would not be captured through student log files.
Unreliability	An unacceptable amount of measurement error invalidates construct inferences.	Database recording errors may threaten reliability. Also, unstandardized conditions of at-scale VLE administration may induce unreliability.[4]	Several of our log files that we aim to use as an indicator of "student usage" contain some unusually long windows of time, such that we are quite sure that database recording errors have occurred, introducing unreliability. In addition, students respond to test items in the platform in an unstandardized manner (e.g., some responses are at night while some are during class time), likely introducing unreliability in test data.
Construct confounding	Inferences are drawn on one construct even though the indicators reflect more than one construct.	Because data collection often occurs before the formal establishment of constructs[2], observed data may often represent multiple constructs.	Our construct of "student individual VLE usage" as measured by computer-logged information is, most likely, also capturing information about access to technology at home.

Confounding constructs with facets of constructs	Only some facets of a construct are measured, invalidating inferences about the full construct.	System logs do not reflect any engagement with the VLE material that takes place outside of the system, such as any written notes, downloaded/printed material, or face-to-face group discussions while engaging with the VLE.	The VLE in our project allows for users to download worksheets related to the VLE videos and practice test items. For participants who download this material, we lose the ability to track their usage as an indicator of any construct of interest. For example, when studying the effect of the VLE on student academic outcomes, the use of printed materials threatens the “treatment dosage” construct, as we are missing some of the facets of exposure to the intervention.
Treatment diffusion	Portions of the intervention (treatment) are diffused to participants who are assumed to not have access to the intervention, threatening the intervention (treatment) construct itself.	Researchers using operational, at-scale VLE data are often not in control of participant usage in the same manner that is typical of traditional experimental research[5], which can result in behaviors incongruent with the assumed intervention behaviors.	We entered the early stages of our VLE research assuming that a student would independently login before engaging with the VLE (i.e., being exposed to the intervention). However, teacher surveys indicated that students work together in the VLE in full classroom groups and in small group settings while only one of the students or teachers logs in. Hence, there are many instances in which we incorrectly assumed that a student who did not login on a particular occasion had zero “treatment dosage” during that occasion. In many cases, the intervention was diffused to those students in a way that the VLE, and hence our log files, did not capture.
Reactivity to the research situation	Participant perceptions of the research project become a part of the measurements, thereby affecting constructs inferred from such measurements.	Some gaming of the system is expected[6] when VLE participants have knowledge (or perceived knowledge) of how the data may be used by researchers or other relevant persons.	According to user survey data, student users of our VLE often know that the teacher will monitor their usage, and we hypothesize that when a VLE video is assigned as coursework, at least some of the students will click on the video and allow it to run, regardless of whether or not they are watching it.

[1] Adapted from Shadish et. al. (2002)

[2] U. S. Department of Education Office of Educational Technology (2012)

[3] Baker (2016)

[4] American Educational Research Association et al. (2014)

[5] U.S. Department of Education Office of Educational Technology (2013)

[6] Baker et al. (2008)

Developing a construct validity table within our project has been helpful for locating identifiable problems that must be addressed to increase validity evidence for effectiveness inferences in our studies. For example, we are currently implementing several solutions to the mono-operation bias example in the first row of Table 1. First, we have evaluated external (i.e., convergent, criterion-related) validity through semi-partial correlations of test item responses to student-level standardized Algebra test scores provided by (State). Hence, we can now administer items that display a high level of external validity, helping to ensure that the relatively small number of indicators of the algebra construct demonstrate a moderate to strong relationship to a more formal construct of algebra ability. Second, we are currently building a multidimensional measurement model within our VLE that makes use of adaptive test item administration of items and constant updating of latent trait estimates, while bearing in mind that such measurement systems must be able to operate at-scale in the VLE (e.g., they cannot introduce currently intractable computational burdens [Park et al., 2018]).

Example methods of addressing other construct validity threats in our project include, but are not limited to (a) implementing random administration of student affect and engagement survey items (Hutt, Grafsgaard, & D’Mello, 2019) as well as periodic teacher surveys about usage patterns, (b) implementing natural language processing methods similar to those in Ozuru et al (2008) for understanding more about the VLE content from which we extract numeric data, and (c) working with teachers and the VLE-developers to encourage the use of online worksheets rather than downloaded worksheets. In sum, we are making a number of efforts to mitigate construct validity challenges in our project and are beginning to remove some problems from our project’s construct validity threat table. Beginning that process by aligning our project to a theoretical framework of construct validity has provided clarity and organization to the solutiongenerating process. While construct validity is a multifaceted concept that can feel daunting to address, the theoretical framework helps to deconstruct the problem into manageable units upon which we can take action.

Discussion

Our examples serve as a brief illustration of aligning a VLE-based research project to a theoretical framework of construct validity. We demonstrated a process that other researchers can adapt to their own projects in order to formally organize thoughts on construct validity. We did not fully exhaust all potential threats to construct validity, nor do we claim that we have excelled beyond other researchers with respect to providing sufficient validity evidence for our construct inferences. Rather, we have found it useful to map our VLE-based data and desired educational constructs onto the formal threats to construct validity, and we believe other research projects will benefit from doing this as well. In addition, we believe the field as a whole can benefit when a community of researchers regularly reflects on formal challenges to construct validity. For example, it may be that there are some particular formal threats that are consistently difficult to overcome when extracting constructs from at-scale, operational VLE data, or that there are some useful, generalizable methods for overcoming particular threats. We can shine a light on such trends if and when researchers utilize common theoretical frameworks. Also, it may be that such evaluations of construct validity can inform future developments of VLEs.

We would like to briefly mention some themes that emerged as we mapped our project onto formal threats to construct validity and began to develop actionable solutions. First, we have experienced many benefits from working directly with the VLE developers and samples of users of the platform. This aligns with recommendations from the U.S. Department of Education’s Office of Technology (2013), which specifically state:

> Developers of digital learning resources, educational researchers, and educators should collaborate to define problems of practice that can be addressed through digital learning and the associated kinds of evidence that can be collected to measure and inform progress in addressing these problems. (pp. xii)

Our project has found these collaborations invaluable as they have allowed us to refine how data is collected in the VLE, to alter the type of data collected in the VLE, and to explore in-depth issues related to the meaning of collected data points. Notably, we have found that the VLE users themselves are excellent informants of what a particular data point may or may not mean.

Second, while some of our attempts to mitigate construct validity problems make use of statistical data mining techniques, we have strongly considered solutions in our project that require changes to how data is collected and interpreted rather than simply extracting more statistical output from the extant data. We did this because we believe that some issues simply cannot be overcome through statistical manipulation of the data. It has been important in our project, and we believe in other projects as well, to overcome construct validity threats by using a combination of statistical techniques, data collection methods, database merging techniques, and data interpretation reflections.

Third, we have mainly explored solutions that can be implemented while collecting data under non-experimental conditions that are inherent to operational, at-scale VLE research projects (U.S. Department of Education Office of Educational Technology, 2013). However, a ubiquitous issue is that we often want to study effectiveness of VLEs on educational outcomes under these non-experimental data collection designs in which the researchers have little to no control over participant interaction with the VLE. While, of course, this can impact internal validity evidence for causal inference, we have also found it to have a deleterious impact on construct validity. For example, it is difficult to provide validity evidence for a "treatment" construct when the treatments (i.e., various interactions with the VLE) are confounded (presumably) with a host of unobserved variables. When possible, we recommend that researchers makes use of experimental designs in addition to the post-hoc data mining and data weighting that is ubiquitous in VLE-based research, in hopes of promoting further construct validity evidence when estimating the effectiveness of VLEs on critical student outcomes.

Fourth, and as a holistic conclusion to this commentary, we have come to realize that it may be necessary for the educational community to modify traditional construct validity frameworks to address the new realities of technology-enhanced research. For example, data errors that arise from even small programming bugs in VLEs may cause major threats to construct validity above and beyond issues of unreliability. Although data recording errors are not unique to technology-enhanced research and, in some cases, can be less problematic than relying on human reported data (Feng et al., 2014), it may be that the threats such issues pose to construct validity are qualitatively different in VLE-based research. In general, the unstructured nature of data collection in VLEs may lend itself to construct validity threats that were not necessarily flagged within previously developed validity frameworks. A community based effort may be needed to modify or supplement traditional frameworks of construct validity threats, a recommendation similar to that of the U.S. Department of Education Office of Educational Technology (2013) in which researchers are encouraged to revitalize frameworks for establishing validity of causal inferences from research in technology-enhanced environments. Ultimately, such causal inferences require evidence of statistical conclusion validity, internal validity, and external validity, all of which relate to and hinge upon evidence of construct validity.

References

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: Author.

Baker, R.S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education, 26*, 600-614.

Baker, R. S., & Ocumpaugh, J. (2014). Cost-effective, actionable engagement detection at scale. *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 345-346).

Baker, R. S., Ocumpaugh, J. L., & Andres, J. M. A. L. (2018). BROMP quantitative field observations: A review. In R. Feldman's (Ed.), *Learning Science: Theory, Research, and Practice.* New York, NY: McGraw-Hill.

Baker, R. S., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research, 19,* 185-224.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review of future and visions. *Journal of Educational Data Mining, 1,* 3-17.

Cope, B., & Kalantzis, M. (2016). Big Data comes to school: Implications for learning, assessment, and research. *AERA Open, 2,* 1-19.

D'Mello, S., Dieterle, E., & Duckworth, A. (2017). Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational Psychologist, 52,* 104-123.

D'Mello, S., Kappas, A., & Gratch, J. (2017). The affective computing approach to affect measurement. *Emotions Review, 10,* 174-183.

Daniel, B. (2015). Big data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology, 46,* 904-920.

DiCerbo, K. E., & Behrens, J. T. (2014). *Impacts of the digital ocean on education.* London, UK: Pearson.

Feng, M., Roschelle, J., Heffernan, N., Fairman, J., & Murphy, R. (2014). Implementation of an intelligent tutoring system for online homework support in an efficacy trial. In *Intelligent Tutoring Systems* (pp. 561–566). Cham: Springer International Publishing. doi:10.1007/978-3-319-07221-0_71

Garcia-Alvareza, M. T., Novo-Cortib, I., & Varela-Candamioc, L. (2018). The effects of social networks on the assessment of virtual learning environments.: A study for the social sciences degrees. *Telematics and Informatics, 35,* 1005-1017.

Hutt, S., Grafsgaard, J., & D'Mello, S. K. (2019). Time to Scale: Generalizable Affect Detection for Tens of Thousands of Students across An Entire Schoolyear *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI 2019)*. New York: ACM.

Idemudia, E.C., & Negash, S. (2012). An empirical investigation of factors that influence anxiety and evaluation in the virtual learning environment. *SAIS 2012 Proceedings, 20*. Retrieved from https://aisel.aisnet.org/.

Kai, S., Almeda, M. V., Baker, R. S., Heffernan, C., & Heffernan, N. (2018). Decision tree modeling of wheel-spinning and productive persistence in skill builders. *Journal of Educational Data Mining, 10,* 36-71.

Klasnia-Milicevic, A., Ivanovic, M., & Budimac, Z. (2017). Data science in education: Big data and learning analytics. *Comput Appl Eng Educ, 25,* 1066-1078.

McNamara, D., Allen, L. K., McCarthy, S., & Balyan, R. (in press). NLP: getting computers to understand discourse. In K. Millis, D. Long, J. Magliano, & K. Wiemer's (Eds.) *Deep learning: Multi-disciplinary approaches.* NY: Routledge.

National Forum on Education Statistics. (2015). *Forum Guide to Elementary/Secondary Virtual Education Data*. (NFES 2016-095). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: The passage or the question? *Behavior Research Methods, 40,* 1001-1015.

Park, J. Y., Joo, S. –H., Cornillie, F., van der Maas, H. L. J., Van den Noortgate, W. (2018). An explanatory item response theory method for alleviating the cold-start problem in adaptive learning environments. *Behavior Research Methods, 51,* 895-909.

Shadish, W.R., Cook, T. D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Belmont, CA: Wadsworth.

Shute, V. J., & Moore, G. R. (2017). Consistency and validity in game-based stealth assessment. In H. Jiao & R. W. Lissitz's (Eds.), *Technology enhanced innovative assessment: Development, modeling and scoring from an interdisciplinary perspective.* Charlotte, NC: Information Age Publisher.

Susnea, E. (2018). How big data analytics will reshape e-Learning. *e-Learning and Software for Education, 4,* 72-76.

U.S. Department of Education Office of Educational Technology. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief.* Washington, DC: Author.

U.S. Department of Education Office of Educational Technology. (2013). *Expanding evidence approaches for learning in a digital world.* Washington, DC: Author.

Weller, M. (2007). *Virtual learning environments: Using, choosing and developing your VLE.* NY: Routledge.