GUEST EDITORIAL PREFACE

Special Issue on Assessing Human Capabilities in Video Games and Simulations

Richard N. Landers, Old Dominion University, Norfolk, VA, USA

ABSTRACT

There is increasing interest in using examination of behaviors during videogame play as measurements of human knowledge, skills, abilities, and other characteristics. Such a shift reflects a growing dissatisfaction with traditional assessment methods, like surveys, tests, and interviews. Game-based assessment addresses this dissatisfaction in three ways. First, test-taking motivation is likely to increase due to the motivating nature of well-designed games. Second, scores may be less contaminated in high-stakes contexts if games are perceived as less threatening than cognitive tests. Third, the validity of measurement may increase because gameplay is a behavioral outcome, whereas other approaches require respondents to engage in accurate self-reflection. Fortunately, the cost of game development has decreased to the point where it is now feasible for individual researchers to develop their own videogames or modify existing videogames in order to test these concepts. Rigorous experimental designs, large sample sizes, a multifaceted approach to validation, and in-depth statistical analyses are recommended, so that the assessment game literature meets the same standards as the validation literature at large, with the long-term goal of replacing many traditional assessments with game-based variations of equal psychometric strength. One day, perhaps "assessment" will be synonymous with "fun."

AN INTRODUCTION TO GAME-BASED ASSESSMENT:

Frameworks for the Measurement of Knowledge, Skills, Abilities and Other Human Characteristics Using Behaviors Observed Within Videogames

The point is not that adequate measurement is 'nice'. It is necessary, crucial, etc. Without it, we have nothing. (Korman, 1974, p. 194)

Accurate assessment of constructs is perhaps the most foundational practice of modern social science. If researchers cannot claim that what they intend to measure is what they are actually measuring, no conclusions drawn from those measurements can be valid. All of social science, whether quantitative or qualitative in nature, falls apart. To make matters worse, designing measures effectively is time-consuming and expensive, fraught with pitfalls and red herrings (Dillman, 2011). A massive research literature on psychometrics, the assessment and analysis of unobservable phenomena using observable indicators, provides the overarching framework by which such judgments can be made. When constructing surveys, a variety of rules and guidelines must be followed for scales to represent what they are intended to represent. When administering an interview, questions must be carefully constructed to precisely target the intended domain without corruption by the researcher's intentions or interests (Moustakas, 1994). Such procedures are among the earliest and most important steps to making legitimate conclusions from research.

Unfortunately, the result of such procedures is often an unpleasant and impractical assessment experience. In the survey context, scale length is a major concern to researchers (Hoerger, 2010). To reliably measure a non-cognitive construct with a survey, anywhere from three to dozens of items may be required per construct, depending upon scale construction, the target domain, and the population to be assessed (Cortina, 1993). Yet research participants, especially those in applied contexts, may not be able to sit down to a several-hundred-item survey. To reliably measure a cognitive construct, like knowledge, skill, or cognitive ability, necessary scale lengths are commonly double or triple those required for non-cognitive measures. In some contexts, tests of such lengths are contaminated by the participant's struggle to concentrate for such a long period of time (Ackerman & Kanfer, 2009), leading to criterion contamination and mismeasurement. In the context of interviews, researchers must go to great lengths to ensure that their own opinions, attitudes, and beliefs do not influence the answers provided by interviewees. Entire research paradigms have been developed precisely to avoid this problem (Moustakas, 1994), which typically results in even longer assessment experiences than when using surveys. Even when measurement conditions and design are optimal, both methods are still limited by the ability of those being assessed to reflect and assess their own capabilities (Furnham, 1986). To the extent that people cannot or choose not to do so, such traditional measurement approaches are inherently flawed.

Assessment via behaviors observed in videogames offers a promising solution to many of these problems; presenting assessment content as a game brings several potential advantages. First, by increasing the intrinsic motivational value of an assessment via game design (Dickey, 2007), test-taking motivation can also potentially be increased, decreasing the effect of prolonged testing periods and increasing the general desirability of assessment. If an assessment game could be made to achieve a similar level of attractiveness as commercial hit videogames while maintaining psychometric properties similar to or better than existing measures, it could be used as a replacement to traditional testing methods. Second, presenting a cognitive test as a game in a high-stakes environment may mitigate some of the effects of test anxiety, which contaminates the validity of test scores when present (Cassady & Johnson, 2002). Third, videogames and simulations ask people to display their capabilities via behavior. Instead of asking people to reflect on themselves, the outcomes of their capabilities can be observed directly. If assessment games can be designed to tap more directly upon these capabilities, the validity of assessment can be increased (Armstrong, Landers & Collmus, *in press*).

Despite these potential advantages, research on these possibilities remains in its infancy. This may be most directly attributable to the traditionally great expense and complexity of videogames. In the past, the creation of any videogame or simulation was prohibitively expensive. It required sizable teams of programmers, artists, designers, and many others. In the modern day, however, videogames can be created by individuals at a relatively low cost. Although they may not have quite the production values of modern commercial videogames, the level of quality possible by a lone developer has increased dramatically over the past decade and is anticipated to continue increasing. The development of a videogame – whether web-based, smartphone-based, consolebased, or anything else – is now within the expense budgets of many individual researchers. Additionally, the increased popularity of mods, in which players augment existing games with additional scenarios, art, or other game assets, allows for assessment games to be created even less expensively. As a result of this change, we have entered what might be considered a golden age of videogame research, in which a vast number of unanswered questions have only recently become relatively easily answerable.

In order to actually conduct such studies, it is important to bring games researchers to the cutting edge of modern psychometric theory. Game studies should not reinvent theory where suitable theory already exists (Landers, Bauer, Callan & Armstrong, 2015), and significant work has already been completed on the topic of measurement. Although introductions to modern quantitative measurement are available for games researchers (e.g., Landers & Bauer, 2015), in-depth treatments are generally lacking.

When creating an assessment game, most foundationally, reliability must be established. Classical measurement theory is focused upon the idea of replication; specifically, by observing a sample multiple times, the mean value obtained across those observations is more likely to reflect a population value. In this way, multiple levels or scenarios of a videogame might be used as an analog to survey scale items. However, it remains unknown just how many levels or scenarios are necessary to obtain a stable estimate of behavior, and it is unclear how levels and scenarios should be designed to minimize the number required. Researchers must also consider learning effects; if a player's performance per scenario increases over time as a result of playing, that performance is unlikely to reflect a long-term underlying human capability. Somewhat antithetical to traditional videogame design, the design of many assessment videogames should be such that it is difficult or impossible to improve performance with practice.

Once reliability is established, test validity is paramount. Reliability is a necessary but insufficient criterion for validity; specifically, a game may reliably measure *something*, but that something may not be the intended construct. Because a measure can never be considered simply "valid" or "invalid" (Landers & Bauer, 2015), the validation of an assessment game involves the compilation of numerous types of evidence from several different types of sources, including evidence from test content, response processes, the internal structure of the measures, and exploration of the nomological net (Messick, 1995). Only after numerous types of validity evidence are available that all converge on the same perspective can an assessment game be said to measure the construct it claims. Holding games to the same research quality standards as other measurement approaches is the first step to their legitimacy as a measurement technique.

To encourage further rigorous research on assessment games, we present four treatments of these concepts in this special issue. In the first two articles, a technical approach is taken examining the psychometric properties of videogame behaviors in relation to existing, validated measures. First, Buford and O'Leary (*this issue*) explore the ability of the critically-acclaimed first-person puzzler *Portal 2* to assess cognitive ability, providing reliability and some convergent validity evidence with fluid intelligence but not with general ability. This suggests that the relationship between gameplay and cognitive ability may be multifaceted and dependent upon gameplay content. Importantly, the researchers did not find a moderating effect of game experience, suggesting game-based assessment can be made equally valid for game players and non-game players. Godwin, Lomas, Koedinger and Fisher (*this issue*) explore the ability of a custom-designed videogame called *Monster Mischief* to assess selective sustained attention, a construct that captures a person's ability to focus and target their effort on a single activity, which is quite important in the context of child learning. In their study, both reliability and convergent validity evidence were presented, alongside evidence that the preschool children within their study found the videogame much preferable as an activity versus the traditional selective sustained attention measure. Across both of these studies, convergence with existing measures was approximately $r = .50$, meaning that 25% of the variance in the existing measure could be explained by the variance in the videogame measure. Although this is promising early evidence, clearly much work remains to be done to increase that value; in the psychometrics literature, the traditional standard for reliability (to ensure multiple measures assess the same construct) is a bare minimum of 70%, although higher values are preferred (Nunnally, 1978).

The last two papers in this special issue focus upon measurement of knowledge. Johnson-Glenberg, Birchfield, Megowan-Romanowicz and Snow (*this issue*) explore the representation of knowledge within two custom-designed games designed for the Microsoft Kinect called *Ratio Match* and *Tour de Force*, comparing emergent gameplay behaviors with a knowledge measure developed for the study. Within their study, they found that the number of switches, a key gameplay behavior, was negatively correlated with performance on that test, suggesting that those who played the game more efficiently did so because they had greater content knowledge. In the final paper, Wang, Shute and Moore (*this issue*) provide a summary of best practices and lessons learned using a more established game-based assessment technique called stealth assessment (Shute, 2011). Within this approach, Bayesian networks are employed to create a system of conditional probabilities associated with individual behaviors within a game. These behaviors can then be used to create a real-time estimate of the player's knowledge, as it increases over the course of a game. The application here is somewhat different from the first three articles in that the purpose of the game is two-fold: players are expected to learn as they play, but accurate assessment is still needed during the learning process. This article provides a fascinating glimpse into the complexities associated with such a task.

In summary, assessment via gameplay behaviors is a young but highly promising area of research. With further development, such assessment techniques could effectively replace the dull, time-consuming, and anxiety-producing traditional approaches commonly used today, including both cognitive and non-cognitive measures in both low-stakes and high-stakes contexts. The dropping costs of videogame development and increasing popularity of near-zero-cost modding enable researchers to explore such assessment in wholly new ways. Critically, during this uptick in research, we must reach for the high standards of research quality typical of other validation research literatures. Rigorous experimental designs, large sample sizes, a multifaceted approach to validation, and in-depth statistical analyses should be the standard, not the exception. With such an approach, perhaps, one day, a person's first thought upon hearing the word "assessment" will be "fun."

Richard N. Landers
Guest Editor
IJGCMS

REFERENCES

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology*, *15*, 163–181. PMID:19586255

Armstrong, M. B., Landers, R. N., & Collmus, A. B. (in press). Gamifying recruitment, selection, training, and performance management: Game-thinking in human resource management. In D. Davis & H. Gangadharbatla (Eds.), *Handbook of Research on Trends in Gamification*. Hershey, PA: Information Science Reference. doi:10.4018/978-1-4666-8651-9.ch007

Buford, C. C., & O'Leary, B. J. (This issue). Assessment of fluid intelligence utilizing a computer simulated game. *International Journal of Gaming and Computer-Mediated Simulations*.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270–295. doi:10.1006/ceps.2001.1094

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and application. *The Journal of Applied Psychology*, *78*(1), 98–104. doi:10.1037/0021-9010.78.1.98

Dickey, M. D. (2007). Game design and learning: A conjectural analysis of how massively multiplayer online role-playing games (MMORPGs) foster intrinsic motivation. *Educational Technology Research and Development*, *55*(3), 253–273. doi:10.1007/s11423-006-9004-7

Dillman, D. A. (2011). *Mail and Internet surveys: The tailored design method*. John Wiley & Sons.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences*, *7*(3), 385–400. doi:10.1016/0191-8869(86)90014-0

Godwin, K. E., Lomas, D., Koedinger, K. R., & Fisher, A. V. (This issue). Monster Mischief: Designing a video game to assess selective sustained attention. *International Journal of Gaming and Computer-Mediated Simulations*.

Hoerger, M. (2010). Participant dropout as a function of survey length in Internet-mediated university studies: Implications for study design and voluntary participation in psychological research. *Cyberpsychology, Behavior, and Social Networking*, *13*(6), 697–700. doi:10.1089/cyber.2009.0445 PMID:21142995

Johnson-Glenberg, M. C., Birchfield, D. A., Megowan-Romanowicz, C., & Snow, E. L. (This issue). If the gear fits, spin it! Embodied education and in-game assessments. *International Journal of Gaming and Computer-Mediated Simulations*.

Korman, A. K. (1974). Contingency approaches to leadership. In J. G. Hunt & L. L. Larson (Eds.), *Contingency approaches to leadership* (pp. 189–195). Carbondale: Southern Illinois University Press.

Landers, R. N., & Bauer, K. N. (2015). Quantitative methods and analyses for the study of players and their behaviour. In P. Lankoski & S. Bjork (Eds.), *Research Methods in Game Studies* (pp. 151–173). Pittsburg, PA: ETC Press.

Landers, R. N., Bauer, K. N., Callan, R. C., & Armstrong, M. B. (2015). Psychological theory and the gamification of learning. In T. Reiners & L. Wood (Eds.), *Gamification in Education and Business* (pp. 165–186). Cham, Switzerland: Springer. doi:10.1007/978-3-319-10208-5_9

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *The American Psychologist*, *50*(9), 741–749. doi:10.1037/0003-066X.50.9.741

Moustakas, C. (1994). Phenomenological research methods. *Sage (Atlanta, Ga.)*.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age Publishers.

Wang, L., Shute, V., & Moore, G. R. (This issue). Lessons learned and best practices of stealth assessment. *International Journal of Gaming and Computer-Mediated Simulations*.