1. Introduction

Serious games are being explored in particular for their potential to provide more valid assessments compared to traditional assessment approaches. It is argued that serious games can provide more meaningful and authentic (i.e., ecologically valid) contexts for assessments through interactive immersive environments (Lumsden, Skinner, Woods, Lawrence, & Munafò, 2016). The increased user engagement in assessments that can be promoted in these contexts through “fun” and interactive game design approaches is further presumed to increase the validity of the assessments compared to typical paper and pencil assessments that can be repetitive and boring (Shute & Rahimi, 2017). In addition, the immersive and interactive environment of serious games is being explored as a means to reduce user test anxiety through a “stealth” approach (Shute et al. 2016). This approach aims to measure performance by unobtrusively logging user behaviors (e.g., mouse clicks on objects indicating choices or strategies, time to complete tasks, number of attempts to complete a mission, paths taken to reach destinations, use of support such as manuals or adjusting difficulty levels) in a computer game's log file rather than explicitly asking a user to self-report their thoughts and behaviors in an evaluative assessment. It is well-documented that assessment methods that rely on non-verbal indices
of certain constructs (e.g., speech fluency/hesitation as an indicator of anxiety) have greater predictive validity compared to self-report measures (Greenwald, Poehlman, Uhlmann, & Banaji, 2009). It is also well documented that anxiety has a negative impact on user performance on tests and in other evaluative situations (e.g., giving a public speech) (Cassady, 2004; Egloff & Schmukle, 2002; Horwitz, 1986). This stealth approach is designed to reduce the “player’s” test anxiety and bias to respond in certain ways (e.g., promote a positive image or give the “right” answer) that are reactive to the more explicit demands of a traditional assessment environment (Shute et al., 2016; Shute, Ventura, Bauer, & Zapata-Rivera, 2009).

Serious games provide a context for measuring and assessing a broader range of skills and constructs compared to traditional assessment approaches. These skills and constructs include competencies that have been identified in various domains as important for success in the “real world.” In healthcare, there is a recognition of the increasing importance of training and assessing constructs such as “non-technical” skills like communication, teamwork, and leadership to promote patient safety and the quality of care (Yule, Flin, Paterson-Brown, & Maran, 2006). In the US education system, there has been an emphasis on impacting 21st Century skills (Rotherham & Willingham, 2010) and using games to assess “real world” competencies such as creativity, critical thinking, and teamwork (Shute & Ventura, 2013). Serious games have perhaps the longest documented history of use in business and management education since the 1950s (Faria, 1987; Wolfe, 1997). They have been used to teach and assess “real world” skills critical to business management such as decision-making, strategic thinking, leadership, and teamwork (Vos, 2015). Serious games are amenable to assessing these competencies because they can be designed to simulate real world environments or take advantage of fictional and fantasy contexts to elicit behaviors that demonstrate creativity, decision-making, teamwork, and leadership that can be assessed through the paths that are taken, how resources are managed and the options chosen. Users’ performance can be tracked, summarized, and then validated against other validated measures or similar behaviors in the real world. Because serious games are collecting information on behaviors in interactive environments, they are hypothesized to be particularly advantageous as a means to measure those behavior-based skills and competencies beyond traditional self-report assessments.

Another advantage of using serious games as assessment tools is their inherent ability to take advantage of current technologies. Serious games in digital (as opposed to non-digital forms such as board or card games) software environments can be programmed to capture, store, and share massive amounts of user data over time through advanced systems and innovative technologies that go beyond traditional “one-shot” assessment measures (de Klerk, Veldkamp, & Eggen, 2015). Sensor technologies that track physiological parameters can also capture massive amounts of user data to provide indicators of behaviors and constructs (Parsons & Reinebold, 2012). This data can be synthesized, analyzed, and fed back to users immediately as an ongoing assessment and/or collected over time to provide insights into user performance profiles for detailed feedback or summative purposes (Rahkila & Karjalainen, 1999).

Building on their technological capabilities to capture user “behaviors,” serious games can be further programmed using artificial intelligence approaches to act as virtual expert assessors of user behaviors and skills. Because they can be programmed to evaluate “behaviors” in a more objective way, they may provide more unbiased, reliable, and precise assessments than “human” test administrators in traditional testing contexts (Bellotti et al., 2013). In addition, because games can provide surrogate “digital” expertise in place of expert “human” assessors, serious games are being explored and evaluated as a means to provide more valid assessments of these “real world” outcomes in ways that could be more cost effective compared to traditional assessment approaches such as pen and paper questionnaires and direct assessment by an expert (Mislevy et al., 2016).

Despite the above outlined advantages, the use of games for assessment is also associated with several challenges. The biggest challenge directly confronts the claims for the advantages of using games for assessment. There is an acknowledged need to evaluate claims that these game design approaches provide more valid assessments (Hernández, Duarte, & Beardo, 2017; Levy et al., 2016; Wiloth, Lemke, Werner, & Hauer, 2016). This challenge is consistent with a broader need in the field of serious games for increased scientific rigor in demonstrating the validity of their use as assessments.
as well as training and intervention tools (Kato, 2012; Levy et al., 2016; Lumsden, Edwards, Lawrence, Coyle, & Munafò, 2016). A recent systematic review (Lumsden et al., 2016) called for increased rigor in evaluating serious game assessment approaches through study designs that increase causal inferences that can be made about the validity of the serious games as assessment tools and increased sample sizes to increase statistical power for increased confidence in concluding that the assessments actually "work". A handful of studies are already engaging in this work by examining associations between user performance in game assessments and traditional validated assessments (Levy et al., 2016; Wiloth et al., 2016).

In addition, the cost-benefits of using games for assessment should also be examined in scientific studies to support arguments claiming this advantage. We were unable to find any published studies showing that using serious games for assessment are more cost effective than traditional assessment approaches. This was not surprising given that there is an acknowledged need for high quality cost-effectiveness studies to support the use of serious games in general in educational settings (Girard, Ecalle, & Magnan, 2013). There is also a need for high quality cost-effectiveness studies with technology solutions in general in broader domains such as health (Shekelle, Morton, & Keeler, 2006; Whitten et al., 2002) and business (Maes, De Haes, & Van Grembergen, 2016). Thus, the need to evaluate serious games as assessment tools for their costs would not only provide evidence for this claim but also contribute to knowledge in other domains.

Another challenge in using games for assessment is the risk that the game mechanics or game environment features used to engage may undermine their validity. For example, typical neuropsychological assessments of Attention Deficit/Hyperactivity Disorder (ADHD) can be boring and difficult to complete for these patients who have difficulties with attention and focus (Ramos-Quiroga, Montoya, Kutzelnigg, Deberdt, & Sobanski, 2013). Although there is often a concern that serious games can be “boring” or less engaging compared to commercial games (Buday, Baranowski, & Thompson, 2012), there is also a risk they can be too engaging. Game design mechanics can lead patients with ADHD to increase their focus and attention in game activities that have been shown to facilitate the release of dopamine in reward and motivation pathways that are typically disrupted in this population (Volkow et al., 2011). In one study (Delisle & Braun, 2011), office tasks that place demands on core deficits of ADHD (e.g., executive functioning skills such as planning and working memory) were simulated in a fast-paced video game environment that was designed to assess core features of ADHD. Users were able to easily carry out these tasks in this environment suggesting that the game features helped the users mask or fail to demonstrate core deficits of ADHD. This suggests that certain game features can undermine the scientific validity of serious games for assessment of specific constructs.

Similarly, serious games for assessment can also raise challenges if their interfaces or engagement mechanics are irrelevant to assessing a construct or criterion (Rupp, Gushta, Mislevy, & Shaffer, 2010). These irrelevant factors, such as objects in the environment, the interface, or colors that increase engagement or create a sense of authenticity, may at the same time introduce error and noise into assessments that can negatively impact the sensitivity and specificity of a game as a tool for assessment (de Klerk et al., 2015; Kraemer, 1992). There is a need for research, such as that conducted recently by Levy and colleagues (2016), that examines trade-offs and “clashes” between elements in game assessment environments intended to promote authenticity or entertainment that may also adversely impact their validity as assessment tools.

## 2. The Contributions to This Special Issue

Overall, the advantages as well as the challenges presented above suggest that there is still much work to be done to develop games for assessment especially regarding scientifically demonstrating their validity. The goal of this special issue is to share the work of experts who have been deeply involved in leveraging the advantages of games for assessment to meet the challenges. Their work exemplifies the state-of-the-art of work in the field. We are particularly pleased that the work includes concrete examples of how interdisciplinary teams can work together to produce engaging and valid games for assessment.

DiCerbo (2017) describes a method inspired by agile project management approaches (Schwaber, 2004) often practiced when developing games. The research efforts take advantage of the structure of the agile development schedule and process to develop and check the validity
of methods to structure and assess user activity stream data as learning constructs in a psychometrically and computationally feasible ways. The process the authors describe provides an example of how the typically “slow” process of research validation can work with and even take advantage of the “fast” process of development to create more valid games for assessment.

Bauer & colleagues (2017) discuss how game structures and design principles provide an adaptive and appropriate context for carrying out formative assessments. The authors present examples of existing commercial games that provide contexts that reflect how instructors conduct formative assessments in traditional classroom activities, such as providing explicit information about goals, collecting evidence of learning and providing feedback to improve future learning. They further discuss how assessments in games can contribute to the broader context of learning in the classroom. The process of developing a formative assessment game focused on argumentation skills is used as an example to demonstrate how an interdisciplinary team of experts can work together to incrementally improve the validity and overall quality of the game based assessment.

3. Conclusions

Serious games can be powerful tools for assessment in diverse fields such as education, health, industry, and the government. Along with their potential to improve the validity of assessments, evaluate behaviorally-based constructs thought to be difficult to assess through self report (e.g., creativity, leadership) and take advantage of advanced technologies for data capture and analysis; they come with challenges as well. In particular, there is a clear need for more research to empirically evaluate claims that serious games can be a means to assess constructs and behaviors with greater validity than traditional approaches. Furthermore, there is a need for research that clarifies which games and what characteristics of these games contribute to or even reduce their validity as assessments. Finally, high quality studies of the cost-effectiveness using serious games for assessment should be conducted comparing their use to traditional assessment approaches.

The experts and their contributions to this special issue on serious games for assessment contribute to these aims by drawing attention to the high quality work that is being done in the area. They raise and discuss scientific research issues while providing practical examples of how games can be developed to high standards for key stakeholders involved in the process of developing games for assessment. They lay the groundwork for rigorously validated and highly engaging games for assessment in the future.

4. Acknowledgements

The authors would like to thank Valerie Shute for her helpful comments on an earlier draft of this article as well as Reid Klion and Cor Sluijter for their substantive feedback and support.

5. References


