Computers & Education 85 (2015) 23–34

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example

Sebastiaan de Klerk [a, b, *], Bernard P. Veldkamp [b], Theo J.H.M. Eggen [c]

[a] eX:plain, Department of Vocational Examination, P.O. Box 1230, 3800 BE, Amersfoort, The Netherlands
[b] University of Twente, Faculty of Behavioral Sciences, Department of Research Methodology, Measurement and Data Analysis, P.O. Box 217, 7500 AE, Enschede, The Netherlands
[c] Cito, P.O. Box 1034, 6801 MG, Arnhem, The Netherlands

ARTICLE INFO

Article history:
Received 7 October 2014
Received in revised form
23 December 2014
Accepted 26 December 2014
Available online 19 February 2015

Keywords:
Evaluation methodologies
Simulations

ABSTRACT

Researchers have shown in multiple studies that simulations and games can be effective and powerful tools for learning and instruction (cf. Mitchell & Savill-Smith, 2004; Kirriemuir & McFarlane, 2004). Most of these studies deploy a traditional pretest-posttest design in which students usually do a paper-based test (pretest) then play the simulation or game and subsequently do a second paper-based test (posttest). Pretest-posttest designs treat the game as a *black box* in which something occurs that influences subsequent performance on the posttest (Buckley, Gobert, Horwitz, & O'Dwyer, 2010). Less research has been done in which game play product data or process data itself are used as indicators of student proficiency in some area. However, the last decade researchers have started focusing on what is happening inside the black box to an increasing extent and the literature on the topic is growing. To our knowledge, no systematic reviews have been published that investigate the psychometric analysis of performance data of simulation-based assessment (SBA) and game-based assessment (GBA). Therefore, in Part I of this article, a systematic review on the psychometric analysis of the performance data of SBA is presented. The main question addressed in this review is: 'What psychometric strategies or models for treating and analyzing performance data from simulations and games are documented in scientific literature?'. Then, in Part II of this article, the findings of our review are further illustrated by presenting an empirical example of the – according to our review – most applied psychometric model for the analysis of the performance data of SBA, which is the Bayesian network. Both the results from Part I and Part II assist future research into the use of simulations and games as assessment instruments.

1. General introduction

The use of computer simulations and games as assessment instruments (from here on referred to as *simulation-based assessment* (SBA)) has increased in popularity in the preceding years. Simulations have already been recognized as powerful learning tools (cf. Mitchell & Savell-Smith, 2004; Kirriemuir & McFarlane, 2004). The general rationale is that SBA has some advantages over traditional paper-and-pencil (P&P) tests and performance-based assessments (PBA) and that it can both expand and strengthen the domain of assessment (Clarke-Midura & Dede, 2010; De Klerk, Eggen, & Veldkamp, 2014). First, from the student's point of view, doing an SBA is more fun and entertaining than doing a paper based test. The storyline driven approach of SBA tends to induce *flow* (Csikszentmihalyi, 1990), which is a psychological state in which people lose perception of time and space. Effectively, students are immersed in the SBA when they experience flow. This may also mean that students are highly motivated and dedicated to completing tasks and attaining goals in the simulation while not being preoccupied by test anxiety (Shute et al., 2010). On the other hand, other students may find it difficult to immerse themselves in a virtual environment, or may get confused with the construct-irrelevant aspects of the simulation (e.g., the interface or specific colors). If so, the use of SBA might have serious implications for some students, especially in high-stakes testing situations. Getting students accustomed to simulations, for example during schooling, is often suggested to overcome these possible negative effects of the use of SBA.

* Corresponding author. eX:plain, Department of Vocational Examination, P.O. Box 1230, 3800 BE Amersfoort, The Netherlands. Tel.: +31 337501005.
E-mail addresses: s.dklerk@ecabo.nl (S. de Klerk), b.p.veldkamp@utwente.nl (B.P. Veldkamp), theo.eggen@cito.nl (T.J.H.M. Eggen).

http://dx.doi.org/10.1016/j.compedu.2014.12.020
0360-1315/

Secondly, SBA provides the possibility to place more emphasis on the application of knowledge in highly contextualized environments rather than the replication of knowledge as is usually the case in P&P tests. For example, through the design and use of interactive tasks in an SBA, WestEd researchers were able to improve measurement of the *conducting inquiry* science practice in middle school (Quellmalz et al., 2013). Other researchers have even started to investigate the possibility to use SBA for very practical professions, for instance medical and security professions (Iseli, Koenig, Lee, & Wainess, 2010; Mislevy, Steinberg, Almond, Russell, Breyer, & Johnson, 2001). With technological possibilities improving on a steady pace, quite possibly the assessment of practical/manual skills or at the least procedural/strategic skills through SBA will become more of a common practice in the future. Again, this specific advantage of SBA might also have a negative counterpart. For learning or formative assessment purposes, a student can develop knowledge, skills, and abilities (KSAs) within a specific, contextualized virtual environment, which means that the KSAs are grounded in deep, specific experiences associated with the environment(s) presented in the simulation. Yet, in a high-stakes testing situation, the use of a contextualized environment induces low generalizability of the students' performance. In fact, an SBA for a summative assessment purpose might best be composed of different modules, based on different contextualized environments and tasks.

Thirdly, SBA offers the possibility to capture student's product data as well as their process data. Product data can be regarded as the final work products that students produce during the SBA, while process data are log file entries that indicate *how* student's produced their work products (Rupp, Nugent, & Nelson, 2012). Process data can be very useful for a formative or diagnostic purpose but they can also serve as a source of evidence for a summative purpose. The amount of process data can become very large as the time spent in the simulation increases. Students interacting with an SBA for some time may produce many pages of process data, which may be interesting to analyze for measurement purposes. The use of process data means that the SBA is no longer treated as a black box, from which a student's proficiency development can only be measured through a pretest-posttest design (Buckley, Gobert, Horwitz, & O'Dwyer, 2010). Of course, not all process data is relevant for the statements that we want to make about a student's proficiency in the construct to be measured. Identifying the elements in the process data that are relevant for measurement and synthesizing and combining those elements with students' product data into a coherent psychometric model reflects one of the major advantages and challenges of using simulations and games as assessment instruments.

1.1. Evidence-centered design

Above, we have discussed some advantages of using SBA and possible negative washback of using SBA in high-stakes testing situations. Another challenge for using SBA lies in defining and specifying coherent and complete psychometric models that fit the data that students' performance in SBA's produces. A useful point of departure for this discussion is the *conceptual assessment framework* (CAF) layer within the *evidence-centered design* framework (ECD) (Mislevy, Almond, & Lukas, 2004). The CAF consists of three separate, though strongly related, models: the *student model*, the *activity model* and the *evidence model*.

The student model relates to what we want to measure, it specifies one or more constructs that we are interested in and want to make statements about regarding students' proficiency. In ECD terms these constructs are called Student Model Variables (SMVs) and they are latent, which means that we cannot directly observe them and have to make inferences about these variables based on the observable variables produced by the performance of students in the SBA (Mislevy et al., 2004). Student models can easily become highly complex in SBA as it is often the case that multiple constructs at the same time define the performance of students in the SBA.

The activity model relates to how an SBA's situations and tasks are designed in which we measure what we want to measure. The activity model consists of all the tasks that are part of the SBA. In SBA, tasks are commonly specified as objectives or goals that students have to achieve during their performance in the simulation. In that sense, tasks in SBA's are often different from traditional item – response question formats that are common practice in traditional tests. In traditional assessment tasks, the activity model variables and values are already known to the assessment developer before the test is presented to the student. For example, a computer-based test consisting of 50 multiple-choice with three alternatives that can all be scored dichotomously (0 = incorrect, 1 = correct), and in which student responses are recorded by mouse clicks.

In general, SBAs have some variables and values of the activity model that are known in advance for every student progressing through the assessment, while others are not. Elements that are known, for example, are the interface or a specific situational feature that is the same for every student. Yet, as students are progressing through the simulation, the simulation may in some cases evolve into different states for different students. In that case, the game condition variables may change, also between students, including the rules, possible actions and interactions that are possible at that specific moment in the SBA. Mislevy et al. (2014) call this the state machine of the SBA. These *dynamic* activity model variables make it more difficult to psychometrically model and interpret a students' performance, because the actual values in the dynamic activity model can only be known and operationalized in the perspective of the state machine.

Building on the activity model, multiple sources of data are recorded and collected during a student's performance in the SBA. For example, reaction times, mouse clicks, navigational paths, or successful completion of objectives. Some, but not all, of these data will function as observable variables (OV) that provide information about SMVs through a measurement model. Which pieces of data can be identified as OV and how these pieces accumulate into a coherent measurement model is specified in the evidence model. The evidence model relates to how we measure what we want to measure. Theory and data are united in the evidence model through two separate, though strongly related processes: *evidence identification* and *evidence accumulation* (Rupp, DiCerbo, et al., 2012; Rupp, Nugent, et al., 2012). The supposed theoretical relationship between SMVs and OVs are formalized in the evidence model on basis of the data that are produced by students performing in the SBA. As mentioned above, simulations and games offer the opportunity to record both product and process data which are subsequently saved into a log file. Process data are usually click-stream data (mouse-clicks, navigation paths, use of tools, etc.) that indicates what action students performed during the SBA to produce product data. In the evidence identification part of the evidence model the product and process data that are relevant for the statements that we wish to make about students' SMVs are identified. There might be multiple steps involved before the specific elements of data in the log files can be considered to take the role of OV. For example, if a simulation also records spoken statements of a student, then these statements first have to be transcribed, analyzed and scored before these performance data elements can be used in a psychometric model. Thus, in some cases, there needs to be a data reduction process from raw log file data to manageable data that can serve as input to the measurement model. Only then, these specific elements of