European Journal of Sustainable Development (2019), **8**, 5, 270-277 ISSN: 2239-5938
Doi: 10.14207/ejsd.2019.v8n5p270

Development of an Integrated Game Based Assessment Approach – The Next Generation of Psychometric Testing

Cătălin Gabriel Ioniță[1], Alina Stanciu[1], Adrian Toșcă[1] and Dan Florin Stănescu[2]

Abstract
Game-based assessment have received a lot of attention in the last decade. In a recent study of human resources practitioners, 75% of participants indicated that they would consider using gamification as part of their own recruitment and selection strategy in the near future. Following the methodological approach previously used in educational environment, two approaches to building and using GBA in the organizational environment can be distinguished: gamified assessment – by gamifying (already existing) psychometric test; psychometric play - use of a game to gather evaluation data. Previous studies highlighted that those applying for a job are eager to use game-based assessment for self-evaluation, especially when these games are available for free. Game-based assessments can also help maintain a high commitment during the evaluation, which reduces the likelihood of some candidates dropping out in the process and also increases the amount of time that data can be collected. Current paper aim at presenting the preliminary efforts made to gamify two psychometric tests, namely spatial and verbal reasoning.

Keywords: Game based assessment; recruitment; spatial reasoning; verbal reasoning.

1. Introduction

During the last decade, the application of game element, game mechanics and game design in non-gaming contexts such as in business, education, and social projects has emerged as a major trend. Gamification, defined as the use of game-play mechanics for non-game applications (Deterding et al., 2011) have become one of the most discussed developments in recent years within the framework of staff assessment, especially in the selection area. Game-based ratings have received a lot of media attention and managed to capture the interests of many organizations (eg Unilever, AXA Group, Deloitte etc.). In a survey of HR practitioners deployed by Cut-e Group in 2017, 75% of participants indicated that they are going to consider gamification as part of their own recruitment and selection strategy in the near future.
Gamification can be used for numerous purposes and in various fields. For instance, gamification has a particularly special place in the HR community. Jacobs (2012), stated that "we can gamify many areas of HR, from talent sourcing through to performance management" (p.14). Modern day HR divisions take an increasingly data-driven approach to people management, i.e., the people analytics approach. Games are a powerful instrument for studying human behavior. In a game, rather than asking someone what they did, you can directly observe their behavior. Games also foster increased participation and motivation, which leads to increased quantity and quality of data. By coupling the data advantage provided by gamification with sophisticated analytic

[1]Structural Management Solutions, Bucharest, Romania
[2]National University of Political Studies and Public Administration, Bucharest, Romania

techniques, meaning can be extracted.

Although the evidence is still insufficient, there are a number of arguments for the benefits of game-based assessment. Overall, employers can use game-based assessment to present an innovative image of the organization and increase their attractiveness to potential candidates without compromising their professionalism. The use of (serious) games as an evaluation tool can extend and even strengthen the field of assessment as this type of games has the potential to reveal both the knowledge and the skills and traits that are more difficult to detect when evaluated through traditional evaluation methods, (De Klerk, Eggen & Veldkamp, 2014; Mislevy et al., 2014).

Game-based assessments can also help maintain a high commitment during the evaluation, which reduces the likelihood of some candidates dropping out in the process and also increases the amount of time that data can be collected (Iseli, Koeig, Lee, & Wainess, 2010; Levy, 2013). Previous studies (Kato & de Klerk, 2017) have also shown that game-based assessment reduces testing anxiety and is more likely to generate genuine responses from candidates because they are immersed in gaming experience (Csikszentmihalyi, 1990).

The use of (serious) games as an evaluation tool can extend and even strengthen the field of assessment as this type of games has the potential to reveal both the knowledge and the skills, and traits that are more difficult to detect when evaluated through traditional evaluation methods, (De Klerk, Veldkamp & Eggen, 2015; Mislevy et al., 2014).

In contrast to a standardized test, which only produces product data, a serious game also provides process data. Process data are mouse clicks, keystrokes, navigational behavior, time stamps etc. (Rupp et al., 2012). Performance in a serious game can produce many pages of log file data in just a short period of time. The challenge is to find meaningful relationships between the data presented in the log files and their relationships to the constructs to be measured in real life.

Even if this effective tool is developing with success especially among big corporations, we need to underline some of its pitfalls. First of all, we need to make some considerations about the costs faced to sustain this new solution. For this type of approach, any organization will need to be supported by experts of gamification and psychologists specialized in psychometrics. It has been estimated that 80% of gamified apps will fail to meet business objectives, primarily due to poor design (http://theundercoverrecruiter.com). In fact, first it is important to understand what the organization is looking for in terms of soft skills, and second, it is essential to translate these needs and requests in the right forms of gamified solutions. Nevertheless, with the psychometric models improving (Mislevy et al., 2014), we might also see game-based assessment being used for a summative or credentialing purpose in the future.

2. Objective

Following the methodological approach already used in educational environment (Al-Azawi et al., 2016), two approaches to building and using GBA in the organizational environment can be distinguished: gamified assessment – by gamifying (already existing) psychometric test; psychometric play - use of a game to gather evaluation data.

Starting from those literature review findings, the objective of this paper is to present the efforts made (psychometric considerations) to gamify two psychometric tests: spatial and verbal reasoning.

3. Requirements of Valid Assessment

One of the most important aspect of any type of assessment is to be valid, accurate and precise. 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 (Landres, 2015). For example, when constructing surveys, a variety of rules and guidelines must be followed for scales to represent what they are intended to represent. Similarly, when administering an interview, questions must be carefully constructed to precisely target the intended domain (Moustakas, 1994).

Similar rules and regulations must be followed also in game based assessment approach. If an assessment game could be made to achieve a similar level of attractiveness as commercial hit games 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).

Although introductions to modern quantitative measurement and psychometric aspects are available for games researchers (Landers & Bauer, 2015), in-depth treatments are generally lacking. When creating an assessment game, most foundationally, reliability and validity must be established. 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, and the internal structure of the measures (Messick, 1995).

Before the data obtained in any assessment activity can be used in psychodiagnostic differential activities, it is necessary to determine whether they meet certain conditions. Since 1967, Lienert has proposed a classification of the main and secondary criteria. Among the main criteria one can find objectivity, fidelity and validity, and among the secondary ones normality, comparability, economy and utility. Bartram (1994) gives almost exclusively attention to fidelity and validity. In Romania, authors such as Schiopu (1997) or Rosca (1972) specify criteria such as standardization, fidelity, validity and sensitivity.

As it can be seen from the literature, there is unanimity in terms of two fundamental criteria, namely fidelity and validity. The fidelity of a test refers to the accuracy with which a test measures a particular feature (Urbina, 2004). This assumes the scores of a test must be reproducible, that is to obtain similar results by repeating the measurement, for the same persons, under the same conditions, with tests measuring the same trait / skill on different occasions (Stan, 2002). Among the best known methods of verifying test fidelity are: test-retest method; the parallel form test method; half-split test method.

The most famous way to test a test's fidelity is to use the test-retest method. This involves administering a test to the same samle of participants in two different rounds.

The correlation resulting from two successive administrations of the same test is called test-retest fidelity index. Practically, the temporal stability of the same test is also measured, which is why this index can also be referred to as the stability coefficient. If the period between the two administrations is relatively low (eg two weeks), this coefficient can also be called a confidence coefficient, indicating the degree of trust that can be given to the instrument used.

The parallel form method assumes either random extraction of samples from a population of items of the same nature, the correlation coefficient obtained indicating the degree of certainty with which a particular trait can be measured, or the use of two different forms of administration of the same items (paper-pencil vs. electronic). The correlation coefficient obtained through the correlation between tests with parallel forms is called the coefficient of equivalence. If the context does not allow the use of parallel forms or the repeated administration of the same test, the split-half test method may be used. This involves creating two sets of items from the original set of items of the test and calculating the correlation coefficient between them.

From a psychometric perspective, Cronbach's alpha is believed to be absolutely necessary, but not enough for a test to be used - this is where the issue of validity become important (Sawilowsky, 2003). Validity is the quality of a test to precisely measure the feature it claims to measure (Stan, 2002). In Legendre's conception (apud Bernier & Pietrulewicz, 1997), validity is the ability of an instrument to really measure what it is to be measured. In practice, we mostly encounter content validity, construct validity, and face validity. Content validity implies accepting the idea that a test is the expression of a sample of items (or tasks) considered by a board of experts to be representative of the measurement of a particular characteristic. In this regard, examining the content validity is based on a detailed examination of the contents of the items in a test and determining the suitability with the whole test.

The construct validity or the theoretical validity is defined as an indication of the degree to which the test measures a specific construct (Stan, 2002). Assessment specialists make predictions about the behavior intended to be tested based on a particular theory, thus making a translation of theoretical variables into observable and measurable behaviors.

The face validity is the one evaluated by profans (users) who appreciate the content of a test to see if it is appropriate to the trait it claims to measure. Because it is a rather vague indicator for test validity, and because of the inherent subjectivity of those requested to evaluate it, it is usually used only in the early stages of building or validating a tests. It can be said that a test has face validity when there is a logical and obvious correspondence between test items and what a test is intended to measure (Stan, 2002). Although this is not an indication of the psychometric validity of a test, it is still a desirable feature because it facilitates the acceptance and involvement of participants in the test activity.

4. Results

Taking into consideration the above-mentioned aspects, in the following we present the analysis of the most important statistical indicators for the original and gamified versions of the spatial reasoning test (cubes) and the verbal reasoning test (propositions).

The spatial reasoning test evaluates the ability to understand complex plans and forms, as well as the ability to manipulate certain forms of two or three dimensions and to identify patterns or relationships between them. This also involves the mental manipulation of spatial forms. The test selected to be gamified involves the identification and counting of a series of cubes distributed in different types of shapes, including those which are not directly visible (for example, behind the front row of the cubes). The paper-pencil version of the test is illustrated in Figure 1. It contains 15 images involving the use of two main abilities, namely the ability to visualize spatial images and spatial reasoning, namely the ability to manipulate and to think mentally with these images.
The Cronbach's alpha value (table 1) for paper-pencil version of the scale (α = .830), is well above the recommended value of .07 (Kline, 2000).

Table 1. Reliability statistics paper-pencil

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.830	.831	15

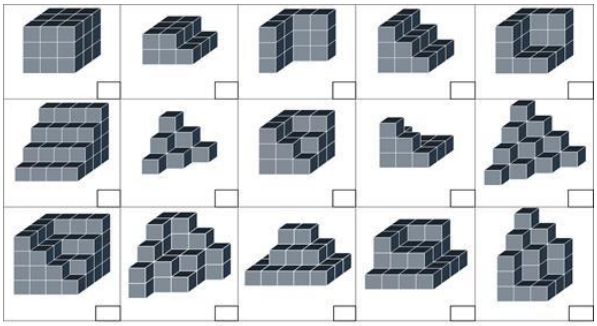

Figure 1: Spatial reasoning paper-pencil version

However, the situation is slightly different in the case of the electronic / gamified version (Figure 2). The gamified version involves running of the 15 items screens in order, the person being evaluated switching from one item to the next one as it provides a response to the previous item.

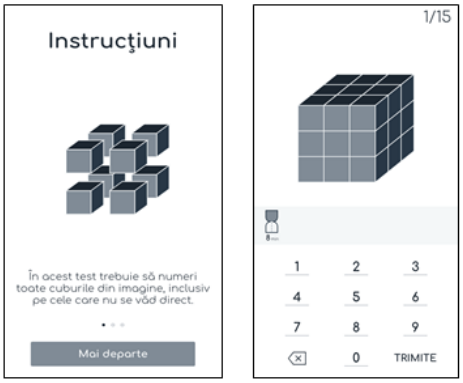

Figure 2: Spatial reasoning gamified version (screen capture examples)

Similarly with the original paper-pencil version of the test, the value of the fidelity index (Cronbach's alpha) for the gamified/electronic version is also above the recommended value (.07) $\alpha = .787$, as can be seen from Table 2.

Table 2. Reliability statistics gamified version

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.787	.797	15

Continuing the fidelity analysis we notice that the value of the correlation coefficient for alternative forms (Table 3), also called equivalency coefficient (pencil-paper and gamified version) is very high ($r = .524$, $p < .001$). In other words, between the original form of the paper (pencil-paper) and the electronic/gamified one, there is a significant positive correlation with a large Cohen effect size.

Table 3. Pearson Correlation parallel forms

		Gamified version
Paper-pencil	Pearson Correlation	.524**
	Sig. (2-tailed)	.000
	N	50

Although the coefficients calculated so far seem to be sufficient for the psychometric validation, we have also decided to calculate the correlation obtained after two successive measures (gamified version), the so-called test-retest fidelity. The gamified version of the test was applied to the same participants sample with a time span of two to three weeks. The test-retest fidelity index is aslo high, $r = .530$, $p < .001$, signifying a strong positive correlation with a large Cohen effect size, the sample exhibiting very good time stability (Table 4).

Table 4. Pearson Correlation test-retest

		Gamified version t1
Gamified version t0	Pearson Correlation	.530**
	Sig. (2-tailed)	.002
	N	31

The data for the verbal reasoning are still in collecting phase and the analysis will follow the same pattern as the one deployed for the spatial reasoning test.

5. Conclusions

In summary, game based assessment 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. Rigorous experimental designs, large sample sizes, a multifaceted approach to validation, and in-depth statistical analyses should be the standard, not the exception.

6. Acknowledgements

This research was supported by the S.I.R.O. project "INNOVATIVE SOLUTION FOR ONLINE RECRUITMENT", contract no. 36/02.08.2017, MySMIS 2014 code 115926

References

Al-Azawi, R., Al-Faliti, F., & Al-Blushi, M. (2016). Educational Gamification Vs. Game Based Learning: Comparative Study. *International Journal of Innovation, Management and Technology, 7(4),* 132–136.

Bartram, D. (1994). Fidelite et validite, in Beech, John, R., Harding, Leonora, *Tests, mode d'employ. Guide de psychometrie*, Paris: ECPA.

Bernier, J. J. & Pietrulewicz, B. (1997). *La psychometrie*, Montreal: Gaetan Morin Editeur.

Cassady, J. C. & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, 27, 270-295.

Csikszentmihalyi, M. (1990). Flow: *The psychology of optimal experience*. New York: Harper & Row.

Cut-e Group. (2017). *White Paper: Ahead of the game. Best practice in games, gamification and game-based assessment*. Retrieved from: https://www.cut-e.com/online-assessment/gamification-in-recruitment/

De Klerk, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2014). A blending of computer-based assessment and performance-based assessment: Multimedia-Based Performance Assessment (MBPA). The introduction of a new method of assessment in Dutch Vocational Education and Training (VET). *Cadmo*, 22(1), 39-56. https://doi.org/10.3280/ CAD2014-001006.

De Klerk, S., Veldkamp, B. & Eggen, T. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & Education. 85*. 10.1016/j.compedu.2014.12.020.

Deterding, S., Dixon, D., Khaled, R. &Nacke, L. (2011). From game design elements to gamefulness: defining gamification. Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, September 28–30, 2011, Tampere, Finland, ACM, pp. 9–15.

Iseli, M. R., Koenig, A. D., Lee, J. J., & Wainess, R. (2010). *Automated assessment of complex task performance in games and simulations* (CRESST Research rep. No. 775). Los Angeles: National Center for Research on Evaluation, Standards, Student Testing. Retrieved from: http:// www.cse.ucla.edu/products/reports /R775.pdf

Jacobs, P. (2012). Emergence of human techsourcing, *Humanresources, 16(6),* 14-15.

Kato, P. M. & de Klerk, S. (2017). Serious Games for Assessment: Welcome to the Jungle, *Journal of Applied Testing Technology*, *Vol 18(S1),* 1-6.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London: Routledge.

Landers, R. N. (2015). An introduction to game-based assessment: Frameworks for the measurement of knowledge, skills, abilities and other human characteristics using behaviors observed within videogames. *International Journal of Gaming and Computer-Mediation Simulations, 7*(4), iv-viii.

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.

Levy, R. (2013). Psychometric and evidentiary advances, opportunities, and challenges for simulation-based assessment. *Educational Assessment*, *18(3),* 182–207. https:// doi.org/10.1080/10627197.2013.814517.

Lienert, G. (1967). *Testaufbau und Testanalyse,* Weinheim: Beltz Verlag.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist,* 50, 741-749.

Mislevy, R.J., Oranje, A., Bauer, M., von Davier, A.A., Hao, J., Corrigan, S., Hoffman, E., DiCerbo, K., & John, M. (2014). *Psychometric considerations in game-based assessment.* New York, NY: Institute of Play.

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

Rosca, M. (1972). *Metode de psihodiagnostic,* Bucuresti: Editura Didactica si Pedagogica.

Rupp, A. A., Levy, R., DiCerbo, K., Sweet, S. J., Crawford, A. V., Calico, T., Benson, M., Fay, D., Kunze, K. L., Mislevy, R. J., & Behrens, J. T. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining*, *4(1),* 49–110.

Sawilowsky, S. S. (2003). Reliability: Rejoinder to Thompson and Vacha-Haase. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 149–154). Thousand Oaks, CA: Sage.

Schiopu, U. (1997). *Dictionar de psihologie* (coord.), Bucuresti: Editura Babei.

Stan, A. (2002). *Testul psihologic. Evolutie, constructie, aplicatii*, Iasi: Editura Polirom.

Urbina, S. (2004). *Essential of psychological testing,* Hoboken, NJ: John Wiley and Sons Inc.