Psychometric Consideration in Game-Based Assessment: An Example of Verbal Reasoning Game

Dan Florin STĂNESCU
National University of Political Studies and Public Administration
30A Expozitiei Blvd., Sector 1, 012104 Bucharest, Romania
dan.stanescu@comunicare.ro

Cătălin Gabriel IONIȚĂ
Structural Management Solutions
95 Alexandru Ioan Cuza Blvd., 011054 Bucharest, Romania
catalin.ionita@structuralmanagement.ro

Adrian TOȘCA
Structural Management Solutions
95 Alexandru Ioan Cuza Blvd., 011054 Bucharest, Romania
adi.tosca@structuralmanagement.ro

Ana Maria IONIȚĂ
Structural Management Solutions
95 Alexandru Ioan Cuza Blvd., 011054 Bucharest, Romania
anamaria.ionita@structuralmanagement.ro

Abstract. *The game-based assessment has received a lot of attention over the past decade, both from industry and media, and has been able to attract attention to many organizations (e.g. Unilever, AXA Group, Deloitte, etc.). In a recent study on human resources specialists, 75% of participants indicated that they would consider using gamming as part of their own recruitment and selection strategy in the near future. Following the methodological approach previously used in the educational environment, two approaches to the construction and use of GBA in the organizational environment can be distinguished: game-based assessment - by gamifying of an already existing psychometric test, and psychometric play – the use of a game to gather the necessary data for the evaluation process. This paper aims at presenting the preliminary efforts made to "gamify" a well-known psychometric test, namely verbal reasoning. The main objective is to present the minimum psychometrics behind the scene necessary to test the validity of the gamified version of the test. Having this in mind, we aim at presenting alternative forms validity, test-retest validity, face validity etc. While GBAs have increased in popularity in the workplace, the research into the validity and reliability of these measures has not lead to conclusive evidence. Due to the lack of conclusive evidence, it is important that more research is conducted to understand how GBAs can be used in the workplace. It is very clear that the potential of games as evaluation tools can only be achieved if data evaluation methods can be developed in psychometric feasible ways because many of the games there are already on the market are based on scenarios or contexts that at best appear to be irrelevant and at worst confuse the role requirements of potential candidates.*

Keywords: *psychometrics; validity; game-based assessment; verbal reasoning.*

Introduction

Advances in technology and psychometric science open the door for a new vision on assessment – game-based assessment. As Klopfer, Osterweil, and Salen (2009) stated, games provide opportunities to both develop and demonstrate proficiencies in complex interactive situations (Klopfer, Osterweil, & Salen, 2009). Therefore, the application of game elements, 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, Dixon, Khaled, & Nacke, 2011) have become one of the most discussed developments in assessment, especially in the personnel selection area. 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.

Due to the fact that more and more, HR divisions take an increasingly data-driven approach to people management, such as the people analytics approach, and games foster increased participation and motivation, which leads to increased quantity and quality of data (Iseli, Koeig, Lee, & Wainess, 2010; Levy, 2013), game-based assessment become the method of choice for many organizations.

Moreover, 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, Oranje, Bauer, von Davier, Hao, Corrigan, Hoffman, DiCerbo, & John, 2014). But, for this type of assessment approach, any organization will need to be supported by experts of gamification and psychologists specialized in psychometrics. It is vitally 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 (Mislevy, Oranje, Bauer, von Davier, Hao, Corrigan, Hoffman, DiCerbo, & John, 2014).

Psychometric aspects

One of the most important aspects 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). Although introductions to modern quantitative measurement and psychometric aspects are available for game 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 one’s normality, comparability, economy and utility. Bartram (1994) gives almost exclusively attention to fidelity and validity. In the Romanian cultural context, authors such as Schiopu (1997) or Rosca (1972) specify criteria such as standardization, fidelity, validity and sensitivity.

The literature review revealed a unanimity regarding 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 alternate/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 sample of participants in two different rounds. The correlation resulting from two successive administrations of the same test is called the test-retest fidelity index (Urbina, 2004, p.124). 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 (e.g. two weeks), this coefficient can also be called a confidence coefficient, indicating the degree of trust that can be given to the instrument used.

Alternate-form reliability procedures are intended to estimate the amount of error in test scores that is attributable to content sampling error. To investigate this kind of reliability, two or more different forms of the test—identical in purpose but differing in specific content—need to be prepared and administered to the same group of subjects (Urbina, 2004, p.126). Thus, 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 or alternate-form reliability coefficient. 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.

One of the most frequently used formulas used to calculate interitem consistency is coefficient alpha (α), also known as Cronbach’s alpha. 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 becomes 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 (Bernier & Pietrulewicz, 1997), validity is the ability of an instrument to really measure what it is to be measured. The view that “test validity concerns what the test measures and how well it does so” (Anastasi & Urbina, 1197, p.113) is still considered as being at the heart of the validity. 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 of 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 refers to the superficial appearance of what a test measures from the perspective of a test taker or any other naive observer 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 nevertheless a desirable feature of tests because it promotes rapport and acceptance of testing and test results on the part of test-takers (Urbina, 2004, p.168).

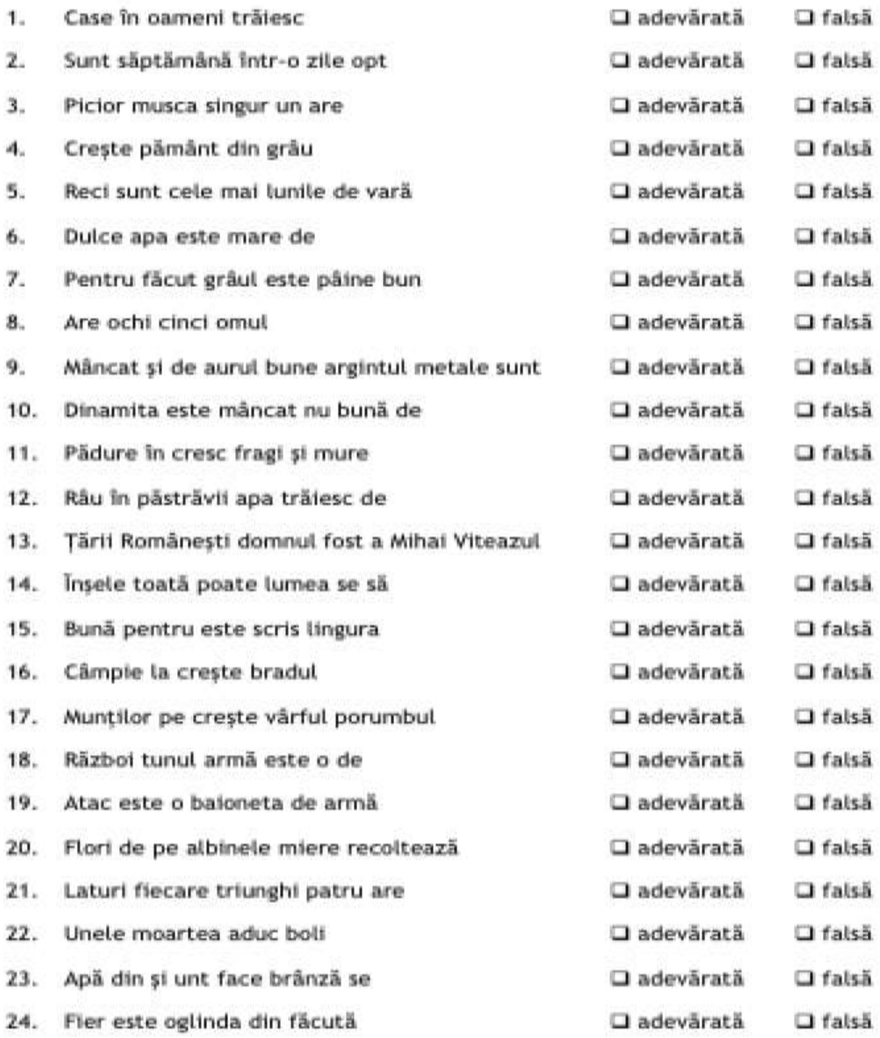
Research objective

As mentioned by Al-Azawi and colleagues (2016), two approaches in 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. (Al-Azawi, Al-Faliti, & Al-Blushi, 2016). The current paper aims at presenting the preliminary efforts made to gamify the verbal reasoning psychometric test. The verbal reasoning test implies not only the understanding of written language and the use of verbal reasoning but also the ability to understand, logically interpret and evaluate written information, rather than just vocabulary recognition or fluency.

Results

Taking into consideration the statistical features previously presented, in the following, we present the analysis of the most important statistical indicators for the original and gamified versions of the verbal reasoning test (propositions). This specific test involves the quick reading and understanding of a series of words presented in a random order, words with which one can compose a meaningful sentence, a sentence whose truth value must be evaluated. For example, from the series of words " have horses feathers all " the sentence "all horses have feathers" can be constructed, a sentence whose value of truth is false.

From a database of 72 items, a series of 24 items will be randomly extracted and the evaluated person will have to provide an answer to each of the items. In Figure 1 you can see the 24 items (Romanian language) selected for the validation tests (alternative forms and test-retest).



1. Case în oameni trăiesc □ adevărată □ falsă
2. Sunt săptămână într-o zile opt □ adevărată □ falsă
3. Picior musca singur un are □ adevărată □ falsă
4. Creşte pământ din grâu □ adevărată □ falsă
5. Reci sunt cele mai lunile de vară □ adevărată □ falsă
6. Dulce apa este mare de □ adevărată □ falsă
7. Pentru făcut grâul este pâine bun □ adevărată □ falsă
8. Are ochi cinci omul □ adevărată □ falsă
9. Mâncat şi de aurul bune argintul metale sunt □ adevărată □ falsă
10. Dinamita este mâncat nu bună de □ adevărată □ falsă
11. Pădure în cresc fragi şi mure □ adevărată □ falsă
12. Râu în păstrăvii apa trăiesc de □ adevărată □ falsă
13. Ţării Româneşti domnul fost a Mihai Viteazul □ adevărată □ falsă
14. Înşele toată poate lumea se să □ adevărată □ falsă
15. Bună pentru este scris lingura □ adevărată □ falsă
16. Câmpie la creşte bradul □ adevărată □ falsă
17. Munţilor pe creşte vârful porumbul □ adevărată □ falsă
18. Război tunul armă este o de □ adevărată □ falsă
19. Atac este o baioneta de armă □ adevărată □ falsă
20. Flori de pe albinele miere recoltează □ adevărată □ falsă
21. Laturi fiecare triunghi patru are □ adevărată □ falsă
22. Unele moartea aduc boli □ adevărată □ falsă
23. Apă din şi unt face brânză se □ adevărată □ falsă
24. Fier este oglinda din făcută □ adevărată □ falsă

Figure 1. *Verbal reasoning paper-pencil version (in Romanian)*

The Cronbach's alpha value (Table 1) for a paper-pencil version of the scale (α = .846), is well above the recommended value of .07 (Kline, 2000).

Table 1. *Reliability statistics paper-pencil*

Cronbach's Alpha	N of Items
.846	21

Continuing the analysis, table 2 presents the contribution of each item of the sample to the scale composite score, as well as the changes of fidelity index value in case of the elimination of certain items. Due to the fact that no variance was observed for 3 items, they were excluded from the analysis.

Table 2. *Item – Total Statistics (paper-pencil version)*

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
VAR00002	14.73	14.601	.104	.848
VAR00003	14.73	15.001	-.134	.854
VAR00005	14.76	14.239	.258	.845
VAR00006	14.73	14.601	.104	.848
VAR00007	14.76	14.039	.361	.842
VAR00009	14.76	14.539	.106	.849
VAR00010	14.83	14.895	-.073	.858
VAR00011	14.71	14.662	.113	.847
VAR00012	14.76	14.939	-.092	.854
VAR00013	14.73	14.751	.014	.850
VAR00014	14.90	13.290	.444	.838
VAR00015	14.78	13.726	.452	.838
VAR00016	14.88	12.860	.626	.830
VAR00017	14.90	12.840	.601	.831
VAR00018	15.05	11.898	.799	.819
VAR00019	15.17	11.795	.797	.818
VAR00020	15.46	13.455	.388	.841
VAR00021	15.22	11.776	.806	.818
VAR00022	15.20	11.561	.873	.813
VAR00023	15.34	12.680	.559	.832
VAR00024	15.27	12.351	.634	.828

However, the situation is slightly different in the case of the electronic/gamified version (Figure 2). The gamified version involves running of the 24 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.

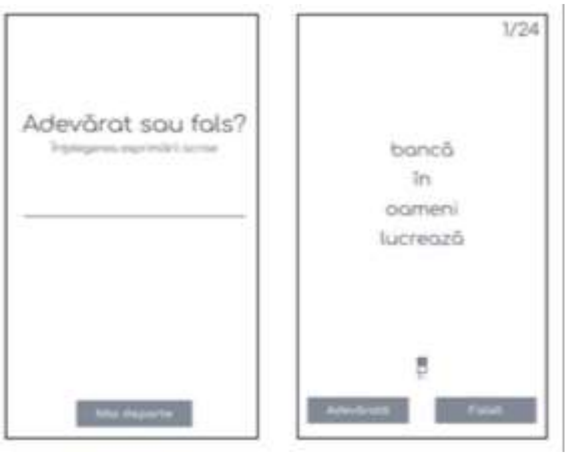


Figure 2. *Verbal reasoning gamified version (sample screens – in Romanian)*

In this case, we must keep in mind that the fidelity index will vary continuously depending on the items randomly extracted from the 72 items in the database. However, for this extraction, the value of the Cronbach's alpha (α = .558) is slightly below the recommended value (.07), as can be seen from table 3. This relatively low value may be a potential problem, but the random extraction of 24 items from the 72 existing ones makes it impossible to calculate all possible extraction variants.

Table 3. Reliability statistics gamified version

Cronbach's Alpha	N of Items
.558	18

The item level analysis (Table 4) reveals an insignificant contribution of certain items to the total score, but their elimination is not recommended considering the constant variation of the items extracted from the 72-item database. Similarly, with the paper-pencil version of the test, we can observe a lack of variation for 6 items, therefore they were excluded from the analysis.

Table 4. *Item – Total Statistics (gamified version)*

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
VAR00001	15.5814	2.583	.154	.550
VAR00002	15.5814	2.583	.154	.550
VAR00003	15.6279	2.525	.109	.556
VAR00004	15.6047	2.483	.228	.539
VAR00005	15.6047	2.578	.084	.558
VAR00006	15.6512	2.423	.188	.544
VAR00007	15.6047	2.483	.228	.539
VAR00009	15.5814	2.725	-.134	.577
VAR00010	15.6977	2.740	-.156	.618
VAR00013	15.6279	2.620	-.007	.575
VAR00015	15.5814	2.725	-.134	.577
VAR00018	15.5814	2.725	-.134	.577
VAR00019	15.5814	2.535	.253	.540
VAR00020	15.6279	2.573	.051	.566
VAR00021	15.6512	2.423	.188	.544
VAR00022	15.6977	1.978	.588	.447
VAR00023	15.8372	1.759	.595	.420
VAR00024	15.7674	1.707	.749	.378

Continuing the fidelity analysis, we notice that the value of the correlation coefficient for alternative forms (Table 5), also called equivalency coefficient (pencil-paper and gamified version) is very high (r = .412, p <.001). Thus, between the original form of the paper (pencil-paper) and the electronic/gamified one, there is a significant positive correlation with a medium Cohen effect size.

Table 5. *Pearson Correlation parallel forms*

		Gamified version
Paper-pencil	Pearson Correlation	.412**
	Sig. (2-tailed)	.005
	N	45

Moreover, the standard test-retest analysis, the correlation calculated from two successive administrations of the test (gamified version) at an interval between two and three weeks, showed a significant positive correlation with a test-retest fidelity index of r = .478, p <. 005, having a medium effect size (Cohen effect size), the sample showing good temporal stability (Table 6).

Table 6. *Pearson Correlation test-retest*

		Gamified version
Paper-pencil	Pearson Correlation	.478*
	Sig. (2-tailed)	.012
	N	27

Also, the face validity, calculated from feedback questionnaires data, showed that 53% of the participants considered that the game is measuring verbal intelligence, verbal, comprehension, language skills, understanding of written expression, verbal fluency, while 47% consider that the game evaluated logical thinking, ability to concentrate and attention.

Conclusions

Although game-based assessment is a young and highly promising area of research, there are several limitations of GBA that will need to be addressed in future studies. The first issue concerns the distinction between GBA and simulation-based assessment. Secondly, it is not yet clear what are the best statistical tools and analyses to be used to collect and process GBA data due to the fact that processing massive and complex gameplay data is difficult and in specific cases time consuming (Leighton & Chu, 2016; Nelson, Erlandson, & Denham, 2011).

The increased usage of GBAs in the workforce increases the need for evidence that these new methods are valid and appropriate for such uses. Even though there is no clear evidence of validity, the research in this respect has fallen behind the adoption of such assessment methods in an organizational environment (Chamorro-Premuzic, Winsborough, Sherman, & Hogan, 2016; Kim & Shute, 2015; Lowman, 2016). With further development, employing rigorous experimental designs, large sample sizes, a multifaceted approach to validation, and in-depth statistical analyses, GBA may represent a great shift in the assessment.

Acknowledgments: *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.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.), Upper Saddle River, NJ: Prentice Hall/Pearson.

Bartram, D. (1994). Fidelite et validite. In Beech, J.R., & Harding, L. (eds.), *Tests, mode d'employ. Guide de psychometrie*, Paris, FR: ECPA.

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

Chamorro-Premuzic, T., Winsborough, D., Sherman, R.A., & Hogan, R. (2016). New talent signals: Shiny new objects or a brave new world?. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 9*(3), 621-640.

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. doi: 10.3280/ CAD2014-001006.

De Klerk, S., & Kato, P.M. (2017). The Future Value of Serious Games for Assessment: Where Do We Go Now?. *Journal of Applied Testing Technology, 18*(S1), 32-37.

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

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, CA: National Center for Research on Evaluation, Standards, Student Testing. Retrieved from http:// www.cse.ucla.edu/products/reports /R775.pdf.

Kim, Y., & Shute, V. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education, 87*, 340-356.

Klopfer, E., Osterweil, S., & Salen, S. (2009). *Moving learning games forward*. Cambridge, MA: The Education Arcade: Massachusetts Institute of Technology. Retrieved from http://education.mit.edu/papers/ MovingLearningGamesForward_EdArcade.pdf.

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.

Leighton, J.P., & Chu, M-W. (2016). First among equals: Hybridization of cognitive diagnostic assessment and evidence-centered game design. *International Journal of Testing, 16*(2), 164–180.

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

Lowman, G.H. (2016). Moving beyond identification: Using gamification to attract and retain talent. *Industrial and Organizational Psychology, 9*(3), 677-682. doi: jpllnet.sfsu.edu/l 0.1017/iop.2016.70.

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.

Nelson, B.C., Erlandson, B., & Denham, A. (2011). Global channels of evidence for learning and assessment in complex game environments. *British Journal of Educational Technology, 42*(1), 88–100.

Rosca, M. (1972). *Metode de psihodiagnostic [Methods of psychodiagnosis],* Bucharest, RO: Didactica & Pedagogica Publishing House.

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 [Psychology Dictionary]* (ed.), Bucharest, RO: Babei Publishing House.

Stan, A. (2002). *Testul psihologic. Evolutie, constructie, aplicatii [Psychological test. Evolution, construction, applications],* Iasi, RO: Polirom.

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