How far has authentic assessment gone? Developments through high-stakes and second language assessment

¿Qué tan lejos ha llegado la evaluación auténtica? Desarrollos a través de la evaluación de segunda lengua y la evaluación con consecuencias importantes

Julio César Gómez*
Universidad Externado de Colombia

Recibido: 18-07-2014 / Aceptado: 12-12-2014

Abstract

Authentic assessment has received considerable attention as a form of assessment that values the learners´ use of skills and knowledge to construct new knowledge through tasks that are meaningful and motivating and still depict features of real-life scenarios. This paper reviews efforts to implement this type of assessment as a substitute for multiple-choice standardized testing while also highlighting its challenges in areas such as validity and reliability that prevented this type of assessment from accomplishing this goal. However, the impact of authentic assessment has gradually made its way into classrooms supported by the well-defined frameworks and outcomes of research. Second language instruction and assessment is one area in which this type of assessment has gained interest and further developments have been put forward.

Key words: authentic assessment, performance assessment, high-stakes testing, second language assessment

Resumen

La evaluación auténtica ha recibido mucha atención como una forma de evaluar las habilidades de los estudiantes para construir conocimiento a través de tareas significativas y motivantes y que, además de ello, simulan escenarios de la vida real. Este artículo revisa los esfuerzos para implementar este tipo de evaluación como un sustituto de las pruebas estandarizadas de elección múltiple y, además, resalta sus retos en áreas como la validez y la confiabilidad, que no permiten que este tipo de evaluación cumpla con su objetivo. Sin embargo, el impacto de la evaluación auténtica ha ido ganando terreno en las aulas con el apoyo de resultados y marcos de referencias de investigaciones. La instrucción y evaluación de una segunda lengua es un área en la que este tipo de evaluación ha ganado interés y ya se han presentado nuevos aportes.

Palabras clave: Evaluación auténtica, evaluación de desempeño, evaluación con consecuencias importantes, evaluación de una segunda lengua.

* Professor of research and academic writing at Universidad Externado de Colombia and temporary member of the graduate faculty at the University of Alabama. He earned his Ph.D. from the University of Alabama. E-mail: jucego69@yahoo.com

Introduction

Authentic assessment has received several names (performance assessment, performance-based assessment, alternative assessment, direct assessment) in testing and educational literature. Despite the difficulty of pointing out the specific nuances of each of these names due to the variety of meanings given and differences in terms of conceptions of knowledge, assessment and learning (Palm, 2008), all of them come to describe those activities in which learners demonstrate all that they know and can do through complex and meaningful tasks (Montgomery, 2002). In the last two decades, authentic assessment has received a great deal of attention and undergone some important developments as a result of the reactions against standardized tests (Worthen, 1993) in the American educational system. In this paper, I will highlight some important characteristics and types of authentic assessments to later describe the challenges advocates of authentic assessment faced in their attempt to establish this type of assessment as a valid form of high-stakes assessment. Then, I will explore models that were designed to address these challenges, some connections of this type of assessment with language testing and the final outcome of this venture in an accountability driven system. I will conclude presenting some general considerations to continue the work with authentic assessment in low-stakes settings and two examples of efforts to support the implementation of authentic assessments in second language teaching and testing.

Background

From the late 80s and the early 1990s, a resurgence of interest in authentic assessment seemed to take over the educational landscape in the United States as teachers became more concerned with the type of multiple-choice standardized tests students had to take (Lewkowicz, 2000). The discreet types of assessment involved more rote memorization of facts and constrained instruction to develop certain kinds of knowledge and test preparation practices. Authentic assessment was regarded as highly appropriate since it focused on the integration of multiple skills and knowledge in tasks (Moss, 1994) that promoted the meaningful performance of the students. Besides, authentic assessment engages learners in the active construction of meaning and doing something with their knowledge (Watson & Robbins, 2008). Once authentic assessment began to be explored, some characteristics were more clearly defined and more elaborated types of tasks designed.

Some authors have identified the most important characteristics of authentic assessment. Brown and Hudson (1998) referring to other authors highlight that authentic assessments require higher-order thinking and problem solving skills, the use of real world situations and simulations and that they focus on both the process and the product. Svinicki (2004), referring to the work of Grant Wiggins, adds the fact that this type of assessment entails judgment and innovation, demonstration of wide range of skills, and the possibility of continuous feedback. Tanner (2001) contributes two characteristics, the first deals with the fact that authentic assessment encourages the transfer of learning to real-life situations beyond the classroom situation and the second features authentic assessment as more responsive to student diversity due to certain degree of flexibility to adjust to their specific characteristics in front of the process and final product. This last aspect is important now that the interactions between the characteristics of the test takers and the characteristics of the tasks are complex (Bachman, 2002) and need to be considered. The focus of the assessment is not only in the final product but in the process and it is carried out through the use of scoring rubrics that demonstrate different levels of proficiency through descriptions of every level of performance (Montgomery, 2002). Closely related to this last characteristic is the

fact that authentic assessments fall in the realm of criterion-referenced testing (O´Malley & Valdez, 1996, Norris, Brown, Hudson & Yoshioka, 1998) since it is based on the definition of criteria for each level of performance upon which responses and products are assessed. These characteristics apply to authentic assessments in all subject areas, however in the next section I will concentrate on how they apply to language testing specifically.

Authentic Assessment and Language Testing

Authentic assessment in language testing somehow follows a similar pattern than assessment in general. The understanding that language is more than the accumulation of skills demanded a form of assessment that could reflect a more holistic and integrative view of language (Hamayan, 1995). Authentic assessment could provide opportunities to engage language learners in authentic situations in which communication and self-expression take place and that could accommodate to the individual learner's needs (Hamayan, 1995). It also can have positive washback effects by providing diagnostic and achievement information in task-based curriculums and aligning classroom assessment with instructional activities (Norris, et al., 1998).

Regarding the types of assessments, some authors have outlined different ways of classifying them. Brown and Hudson (1998) frame authentic assessments in two groups: constructed-response and personal response. Constructed-response assessments require students to produce written and spoken language allowing the integration with receptive skills like reading and listening. In this group, the authors include performance assessments (e.g., essay writing, interviews, problem solving tasks, role playing and group discussions) as well as fill-in and short answer assessments. The personal response also engages students in producing language but the type of response is quite unique for each learner. Examples of assessments in this group include portfolios, conferences and self and peer assessments. The last two assessments introduce the value of the reflection on the process and strategies for learning. Svinicki (2004) offers other two more complex assessments in the form of simulated environments and the creation of real products. Hamayan (1995) introduces learning logs and journals and classroom projects. This list of types of authentic assessments is not exhaustive since the realm of possibilities in each field is quite extensive. Nevertheless, many of the types of assessment presented above have been used broadly in authentic assessment in language testing.

Challenges of Authentic Assessment

The increasing implementation of authentic assessment led educators to identify issues that were going to impact the potential of this type of assessment to become the replacement of multiple-choice standardized test in place. Issues of validity and reliability of authentic assessments impregnated the discussion and motivated study in the assessment and testing field.

Brown and Hudson (1998) point out how the "alternative" nature of these assessments as a "new way of doing things" could lead to believe that they were not to abide to "the requirements of responsible test construction and decision making" (p.567). Linn, Baker, and Dunbar (1991) also highlight the need to establish standards of quality that authentic assessments have to satisfy. It is especially important to do this since many times this performance assessments are used as well to make decisions about students (Norris, et al., 1998) and these standards can contribute to developing sound authentic assessments. Several authors have analyzed these problems and have contributed ways to deal with these issues constructively.

Linn et al. (1991) propose a broader view of the validity of authentic assessments that included

the consideration of criteria such as their consequences, fairness, transfer and generalizability, cognitive complexity, content quality and coverage, meaningfulness and cost and efficiency. In this view of validity we can see how reliability is subscribed in the criterion transfer and generalizability as it is related to variability due to task and rater. Worthen (1993) on the other hand, suggests addressing 12 issues to strengthen the internal rationales of authentic assessment including conceptual clarity, mechanisms for self-criticism, support from well-informed educators, technical quality and truthfulness, standardization of assessment judgments, ability to assess complex thinking skills, use of technology, acceptability to stakeholders, appropriateness to high-stakes assessment, feasibility, continuity and integration across educational systems, and avoidance of monopolies. These last four directly linked to the interest in accommodating authentic assessment to demands of the accountability testing system. Once again, reliability appears reflected in the form of reaching standardization among raters.

Other authors concentrate on the construct to make their argument on the validity of authentic assessments. Miller and Linn (2000) present six aspects of construct validation that are intended to guide the design and evaluation of authentic assessments. The six aspects are content (the task and the construct it is intended to measure), substantive (i.e., examination of the processes used by the learners), structural (i.e., the scoring system and the construct being measured), generalizability (the consistency and replicability of assessment results across tasks and scorers), external (presence of other constructs), and consequential (i.e., intended and unintended consequences and interpretation of scores). This framework includes reliability in two of its aspects: generalizability and structural. It is in this last aspect in which they offer models to augment scorer reliability such as multiple readers, training, benchmarks and calibration checks for drift.

A tendency to approach reliability as part of the validity construct is clear as reflected by the work of the authors presented above. However, one author, Moss (1994) concentrates on reliability across tasks and readers as a factor that in itself requires attention. In order to approach the complexities involved in "reading" or assessing an individual's response to a task, she bases her argument in hermeneutics. Thus, she longs for a more holistic and integrative approach to the interpretation of these individuals´ responses considering all relevant evidence. The role of teachers is underscored as well as that of the students as contributing to the process of assessment. Nevertheless, it is not clear how such approach can be realized in the practical world.

One last approach to address issues of validity of authentic assessments comes from the field of language testing. Bachman (2002) first defines three qualities as essential to any assessment tasks as related to language: construct validity, authenticity and interactiveness. Then, he identifies three elements to be considered in the design, development and use of authentic assessments. The first element is a cognitively-based model of language use and language ability that identifies the characteristics of language use, language ability, topical knowledge, affect and how they all correlate in an assessment task. The second element is a clearly identified domain of target language use situations and tasks that will determine the criteria for the design and development of language tests and the further interpretation of test performance. The third element is a set of distinguishing characteristics for describing assessment tasks and target use tasks including setting, rubric, input, expected response, relationship between input and response. The extent these elements are kept in mind, the more likely authentic assessment tasks may reach high levels of validity. Reliability is not directly considered in this last approach.

This overview of frameworks and approaches to deal with issues of validity and reliability of authentic assessments point out an interest in creating a great deal of awareness of those factors that may directly have an impact on the generalizability of the results of this type of assessments. In the next section, I will explore where authentic assessment stood after more of a decade of experimentation and research in the area.

Outcomes of Research

Supovitz (2009) represents the situation of authentic assessment from the middle of 1990s to 2008 as a "rise and fall" due to the significant increase of attention it received at the beginning of the decade and its later disregard in the test-based accountability system. As stated above, in the early 1990s, authentic assessment got recognition among educators as a more effective way of assessing students' learning in schools in front of criticized existing standardized testing practices. Such was the appeal of authentic assessment that at some point it was considered to become another important form of assessment to go beyond the multiple-choice type of testing. Critics of this type of testing ascertained the serious concerns about gender bias, ethnic prejudice and socioeconomic favoritism it favored which were also confirmed with research. Authentic assessment became the response to get a better measure of student performance that could also impact instruction and curricular reform (Supovitz, 2009). Several states (e.g., Kentucky, Maryland) undertook the implementation of more authentic forms of assessment and research on the growing use of alternative assessment was carried out indicating motivational and alignment trends among administrators and teachers and very positive results of its impact on students.

For instance, a pilot project conducted by Newmann, King and Carmichael (2007) in Iowa aimed at supporting the student production of authentic intellectual work (AIW) by defining a framework to implement authentic assessment and instruction extensively in schools. This framework was based on criteria that included "*the construction of knowledge* through the use of *disciplinary inquiry* to produce discourse, products or performance that have *value beyond school*" (3). After conducting a review of related studies in the United States in grades 3 to 12 including the four major subjects in which authentic intellectual work was observed, these scholars highlighted the higher achievement for students in measures of authentic intellectual performance and standardized basic skills tests. These studies also underscored the positive impact on equal opportunity in education although variability in the levels of authentic instruction at the micro and macro levels makes its impact quite limited. Another important aspect brought up by Newmann et al. (2007) is the degree of involvement that schools and main stakeholders still need to exercise in the development of curriculum, scoring rubrics and in-service training schemes to comprehensibly implement this framework of authentic intellectual work.

Other research has shown how alternative assessment accentuated gender differences, that alternative assessment was cost prohibitive, teachers tended to have a narrow view of the curriculum, and psychometric quality of authentic assessment showed mixed results regarding reliability in scoring. All the problems identified here made it hard for authentic assessment to become a real way of addressing the problems of standardized testing (Supovitz, 2009) and was somehow left aside.

Despite the results of the research on the experimentation with authentic assessment in several schools and its subsequent disregard as an option in front of multiple-choice standardized tests, Supovitz (2009) affirms that traces of authentic assessment have made its way into standardized tests with open-ended writing tasks and some performance tasks.

Authentic assessment still can be considered a very good option of assessment when used in low-stakes environments such as classrooms as a way to meaningfully relate instruction to students' performance (Worthen, 1993). It is also relevant at this level to consider all the frameworks and models presented in this paper to strengthen the validity and reliability of authentic assessments. In addition, it is sensible to consider other strengths of authentic assessments such as the type of feedback and the possibility of moving into more formative views of assessment.

Taras (2005) brings up the idea of formative assessment as directly linked to authentic assessments. However, she indicates that even this type of formative assessment considers the judgment of the performance in relation to a reference level with the only difference being the formative feedback is provided to the learners. This feedback becomes information intended to encourage the modification of the actual level of performance to decrease the discrepancy with this reference level. The clear definition of criteria of assessment gains relevance since it is the point of reference to make judgments and purposefully accompany learners in their process of creation of products.

Similarly, Watson and Robbins (2008) contribute a view of assessment *for* learning directly related to an understanding of teaching to promote learning more than a view of assessment *of* learning which would constitute the reporting of grades and scores. It is a vision of assessment as a tool to support students' learning. There are many benefits for students as they develop more active and reflective roles in the process of elaboration of projects or performance in tasks that can lead to building support relationships and empowerment.

The complexity of the process of implementation of authentic assessment is highly related to the environments in which it takes place. In the United States authentic assessment went through a process that involved high expectations for a solution based on the need to find alternative ways of assessing students' learning in a test-based accountability system and ended up becoming an alternative at the classroom level. Other contexts may experience the implementation of this type of assessment in different ways and they need to be researched extensively.

Implementation of Authentic Assessment in English Language Instruction

There have been some efforts to support the implementation of authentic assessment in English language teaching and to design task-based performance test of English language proficiency and two of them are described as follows.

O'Malley and Valdez Pierce's (1996) book *Authentic Assessment for English Language Learners* presents guidelines for the design and procedures to use authentic assessments with English Language Learners (ELL) in English as a second language (ESL) and bilingual programs in language arts and content areas. This book is a valuable tool to support the integration of authentic assessments in classrooms where teachers are interested in improving teaching and learning.

One project to design Assessment Language Performance (ALP) instruments and procedures for language programs was developed by Brown, Hudson, Norris and Bonk (2002). They conducted a study to examine several means to evaluate performance on test tasks that simulated real-life tasks. Four types of criteria (i.e., task-dependent rating scales, overall performance through task-independent rating scales, task difficulty estimations and test takers' self ratings of their own performance abilities) were used to assess and analyze 13 task-based performance assessments to determine similarities and differences in test

outcomes. A total of 90 participants that belonged to both ESL and EFL contexts were included. The results showed high levels of reliability among raters of the performance of the test takers in the test tasks. This may be due to the careful development and administration of the tasks and rating scales. This project represents the most comprehensive attempt to create an authentic assessment test for language testing that has been reported recently.

Conclusions

In the current paper, I attempted to take a glance at authentic assessment through the process it went through in the United States in the last two decades. The development authentic assessment went through in its attempt to comply with the requirements of an accountability-based testing system led to the identification of issues that have gradually been addressed but ultimately failed to replace multiple-choice standardized testing. However, some features of authentic assessment have been adopted not only in high-stakes testing but also as alternative to more traditional types of assessment in classrooms. Areas such as second language testing have made some important advances in the implementation of authentic assessment at the classroom level and even in the design of tests of language proficiency based on authentic assessments following rigorous parameters of development. It is very likely that new efforts to develop authentic assessments in various areas continue to be undertaken and we can just expect them to be documented soon. In the near future, a new trend on assessment may even put authentic assessment back on the spotlight posing new challenges on what we can attain with it and on what we still need to figure out.

References

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurements: Issues and Practice, (21)*3, pp.5-18.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly,* (32)4, 653-675.

Brown, J. D., Hudson, T., Norris, J. & Bonk, W. J. (2002). *An investigation of second language task-based performance assessments*. Honolulu: University of Hawai'i Press.

Hamayan, E. V. (1995). Approaches to alternative assessment. *Annual Review of Applied Linguistics, (15),* pp.212-226.

Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing, (17)*1, pp.43-64.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, (20)*8, 15-21.

Miller, D. M., & Linn, R.L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, (24)*4, pp. 367-378.

Montgomery, K. (2002). Authentic tasks and rubrics: Going beyond traditional assessments in college teaching. *College Teaching, (50)*1, pp.34-39.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, (23)*2, 5-12.

Newmann, F. M., King, M. B., & Carmichael, D. L. (2007). Authentic instruction and assessment. *Des Moines: Iowa Department of Education.*

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Honolulu: University of Hawai'i Press.

O´Malley, J. M., & Valdez Pierce, L. (1996). *Authentic assessment for English language learners: Practical approaches for teachers*. Reading, MA: Addison-Wesley.

Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation, 13*(4), pp.1-11.

Supovitz, J. (2009). Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform. *Journal of Educational Change, (10)*2-3, pp.211-227.

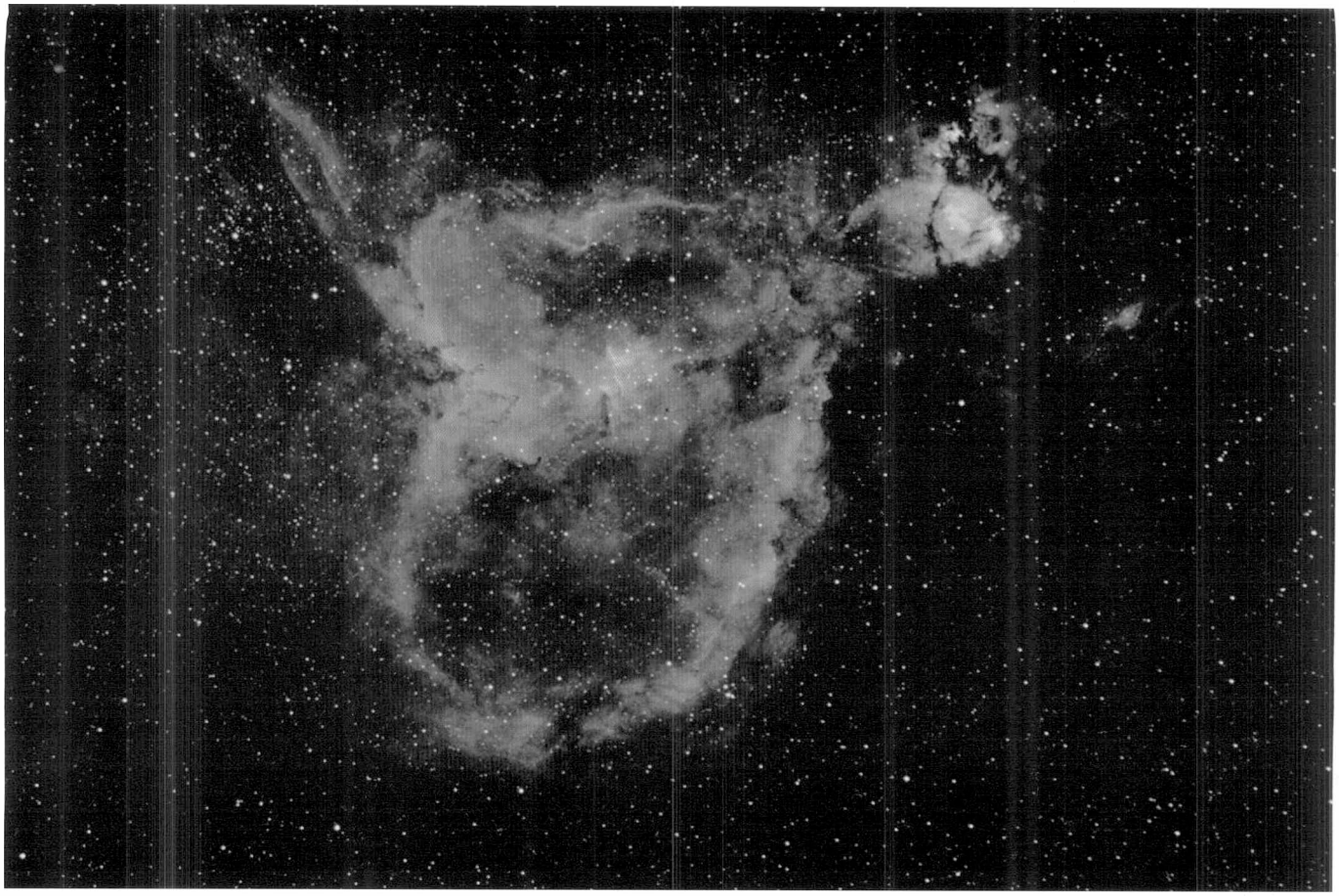
Svinicki, M. D. (2004). Authentic assessment: Testing in reality. *New Directions for Teaching and Learning, (100)*, pp.23-29.

Tanner, D. E. (2001). Authentic assessment: A solution, or part of the problem? *The High School Journal, (85)*1, pp.24-29.

Taras, M. (2005). Assessment- summative and formative- some theoretical reflections. *British Journal of Educational Studies, (53)*4, pp.466-478.

Watson, D., & Robbins, J. (2008). Closing the chasm: Reconciling contemporary understandings of learning with the need to formally assess and accredit learners through the assessment of performance. *Research Papers in Education, (23)*3, pp.315-331.

Worthen, B. R. (1993). Critical issues that will determine the future of alternative assessment. *The Phi Delta Kappan, (74)*6, pp. 444-448, pp. 450-454.



EL CORAZÓN DEL UNIVERSO

Esta bella nebulosa llamada del corazón, por su peculiar forma cuando se toma a exposiciones largas, el objeto IC 1805, en la constelación de Casiopea a 7.500 años luz, una región asombrosa donde nacen de estrellas.

Foto: *Leonardo Delgado Ariza*

Datos: *16 x 600s a través de un filtro de 3 nm Ha OIII y SII, y Canon 200mm f/2.9. Cámara Atik 460EX.y el Telescopio Meade LX200 de 10 pulgadas.*