Gist Education and Learning Research Journal. ISSN 1692-5777.
No.19 (July - December, 2019). pp. 6-30.

Authentic Assessment in the Listening Comprehension Classroom: Benefits and Implications[1]

Evaluación auténtica en el aula de comprensión auditiva: Beneficios e implicaciones

Henry Sevilla Morales and Lindsay Chaves Fernández[2]*
Universidad Nacional de Costa Rica, Costa Rica

Abstract

This research paper discusses the benefits and implications of bringing authentic assessment into listening comprehension classes. The study was run in 2016 based on a mixed-methods model to research and included 38 college students enrolled in a listening comprehension class at an English Teaching Major (ETM) from the University of Costa Rica (UCR). Data collection instruments included plans of improvement, portfolios, self-assessment forms, teacher-student conferences, verbal calls, and impromptu reflections. Data were validated through several procedures (e.g., triangulation and reflexivity) and analyzed in the form of emerging themes from the information collected. Findings are that authentic assessment can and should be used more in listening comprehension classes to bring assessment and instruction together, as well as to provide opportunities for skills integration. The study yields implications for theory and practice, and it constitutes a proposal to move from traditional to process evaluation, and from norm-referenced testing towards more criterion-referenced assessment. Nonetheless, the aim should not necessarily be a radical 'no' to paper-and-pencil tests, but a more balanced use in combination with other strategies so that assessment becomes more reliable, valid, fair, and authentic for all EFL actors involved.

Keywords: authentic assessment, testing, listening comprehension, portfolio, weekly plans

[1] Received: October 1st 2018/ Accepted: July 21st 2019

[2] henry.sevilla.morales@una.ac.cr; lindsay.chaves.fernandez@una.ac.cr

Resumen

La presente investigación discute algunas bondades e implicaciones de implementar la evaluación auténtica en clases de comprensión auditiva. El estudio se desarrolló en el año 2016 mediante un modelo de investigación mixta y contó con la participación de 28 estudiantes universitarios matriculados en un curso de comprensión auditiva de una carrera de enseñanza del inglés (CEE) de la Universidad de Costa Rica (UCR). Los instrumentos de recolección de datos incluyeron planes semanales de mejora, portafolios, reuniones entre el profesor y los estudiantes, intervenciones orales y reflexiones improvisadas *in situ*. Los datos se validaron mediante distintos procedimientos (por ejemplo, triangulación y reflexividad) y se analizaron por medio de categorías emergentes de la información recabada. Los hallazgos sugieren que la evaluación auténtica puede y debe ser utilizada con mayor frecuencia en las clases de comprensión auditiva. Esto, a fin de unificar la evaluación y la enseñanza, además de generar espacios para la integración de varias destrezas del idioma. El estudio genera implicaciones a nivel de teoría y práctica; asimismo, sugiere un cambio de paradigma evaluativo, reemplazando la evaluación tradicional por la de proceso, y las pruebas estandarizadas por las referenciales. No obstante, no se plantea un reemplazo radical de los exámenes escritos, sino un balance adecuado, en conjunción con otras estrategias en procura de una evaluación más fiable, más válida, más justa y más auténtica para todos los actores de enseñanza de lenguas extranjeras involucrados.

Palabras clave: evaluación auténtica, medición, comprensión auditiva, portafolio, planes semanales

Resumo

A presente pesquisa discute algumas bondades e implicações de implementar a avaliação autêntica em aulas de compreensão auditiva. O estudo se desenvolveu no ano 2016 por meio de um modelo de pesquisa mista e contou com a participação de 28 estudantes universitários matriculados em um curso de compreensão auditiva de uma carreira de ensino do inglês (CEE) da Universidade da Costa Rica (UCR). Os instrumentos de coleta de dados incluíram planos semanais de melhoria, portfólios, reuniões entre o professor e os estudantes, intervenções orais e reflexões improvisadas *in situ*. Os dados se validaram mediante diferentes procedimentos (por exemplo, triangulação e reflexividade) e se analisaram por meio de categorias emergentes da informação recebida. As descobertas sugerem que a avaliação autêntica pode e deve ser utilizada com maior frequência nas aulas de compreensão auditiva. Isto, com o fim de unificar a avaliação e o ensino, além de gerar espaços para a integração de várias destrezas do idioma. O estudo gera implicações ao nível de teoria e prática; da mesma forma, sugere uma mudança de paradigma avaliativo, substituindo a avaliação tradicional pela de processo, e as provas padronizadas pelas referenciais. No entanto, não se propõe uma substituição radical das provas escritas, senão um balanço adequado, em conjunção com outras estratégias em procura de uma avaliação mais fiável, mais válida, mais justa e mais autêntica para todos os atores de ensino de línguas estrangeiras envolvidos.

Palavras chave: avaliação autêntica, medição, compreensão auditiva, portfólio, planos semanais

Introduction

In the changing worldwide scenario of modern language education, authentic assessment has gained unprecedented attention from various researchers, scholars, and practitioners (Ali & Ajmi, 2013; Brown & Abeywickrama, 2010; Charvade, Jahandar, & Khodabandehlou, 2012; Frey, Schmitt, & Allen, 2012; Li, 2013; and Moya & O'Malley, 1994). Nonetheless, in actual classroom practice, authentic assessment remains, at best, a dilemmatic subject (Moya & O'Malley, 1994). Some of the factors contributing to this include perceived impracticality and unreliability (Brown, 2010), teachers and students' lack of familiarity with this type of evaluation, and the large amounts of time invested and cost of resources (Murphy et al., 2017). In addition to these, scholars have identified related issues including resistance from institutional authorities and stakeholders, lack of willpower and funding, and reported degrees of subjectivity in the way these evaluations are conducted. And as researchers make their way into this relatively-new field of inquiry, challenges continue to emerge, and questions continue to arise for theoreticians, institutional authorities, and teaching practitioners the world over.

At a theoretical level, perhaps the most evident limitation is the lack of empirical studies on authentic listening assessment in English Language Teaching (ELT). Compared to the general popularity of language assessment, research on this remains rather sporadic and unsystematic (see Dewi, 2018; Li, 2013; Miller, 2003; Yurdabakan & Erdogan, 2009). In order to help bridge this knowledge gap, the current study set out to answer the following research question: What are the benefits and implications of bringing authentic assessment into listening comprehension classes in an EFL program at a public university in Costa Rica? Data were collected from 38 junior students from an English teaching program in Costa Rica. Using a mixed-methods design and different data collection instruments, qualitative data were triangulated to cross-check information and satisfy methodological trustworthiness.

Since changes in teaching paradigms should necessarily imply changes in assessment trends (O'Malley & Valdez, 1996), this paper serves a baseline to move from traditional to process evaluation: from standardization to localness, and from norm-referenced testing towards more criterion-referenced assessment. Findings assist our understanding of the benefits and implications of implementing authentic assessment in listening comprehension classes, and they help bridge gaps between theory and practice as they bring theoretical precept and classroom reality together while implementing authentic assessment. Lastly, they bring us closer to the qualitative assessments which many seem to endorse but which few are examining through empirical research.

Theoretical Background

Some Initial Considerations

While the list of publications in the subject area is vast (e.g., Airasian, 2001; Ali & Ajmi, 2013; Brown & Abeywickrama, 2010; Cohen, 1994; Curtis & Nunan, 2001; Douglas, 2000; Gamboa & Sevilla, 2016a; Genesee & Upshur, 1996; Herman, Aschbacher, & Winters, 1992; McKay, 2006; O'Malley & Valdez, 1996; and Rojas, 2004) and we have therefore left some works unexamined, we hope that for the goal of this paper our brief discussion addressees the core issues put forward during the past few decades.

As some authors have made it clear, the notion of authentic assessment is not new and has been subject to changes throughout the years. As early as the mid' 1980s, increased attention was paid to the benefits of self-assessment strategies such as portfolios (Charvade, Jahandar, & Khodabandehlou, 2012), probably inspired by the ways in which many painters, writers, and other artists displayed their professional skills through portfolios (Bailey, Curtis, & Nunan, 2001; O'Malley & Valdez, 1996; Zollman & Jones, 1994, as cited in Charvade, Jahandar, & Khodabandehlou, 2012). From then on, important advancements have been made both in the alternatives available to teachers and, more crucially, in the way assessment is viewed. Brown & Hudson (1998) challenged the popularly held idea of authentic assessment as a form of alternative assessment (see Herman, Aschbacher, & Winters, 1992) and advocated instead for the term *alternatives in assessment* because, as they argued, "why […] should we even refer to the notion of "alternative" when assessment already encompasses such a range of possibilities?" (Brown & Abeywickrama, 2010, p. 123). Tests then came to be viewed as a subset within the wider scope of assessment, so that, to date, a number of key characteristics have been proposed for all types of assessment: usefulness, purpose, reliability, validity, practicality, washback, authenticity, transparency, and security (see Brown and Abeywickrama, 2010; Combee et al., 2007; Gamboa & Sevilla, 2013; Gamboa & Sevilla, 2016a; and Rogier, 2014). These characteristics are roughly summarized below.

Usefulness means the extent to which a test is congruent with the needs of a specific audience; it has been considered by some as the most important testing principle, and it is closely linked to the concept of *purpose* (or test goal) (Bachman and Palmer, 1996, as cited in Rogier, 2014). In order for a test to be useful, its purpose must match the test taker's reasons for learning the language. *Reliability* deals with the consistency of student scores over time; that is, the assessment's ability to render similar scores if administered in a different moment and place to students with similar characteristics (Brown, 2010; Combee et al., 2007; Rogier, 2014). *Validity* is, in essence, an assessment's capacity to assess what it is supposed to assess (Brown, 2010;

Combee et al., 2007;). *Practicality* deals with the assessment's being easy to create, easy to administer, and easy to evaluate (Brown & Abeywickrama, 2010; Combee et al., 2007; Rogier, 2014). *Washback* makes reference to the effect of assessment on the curriculum; in other words, it is the extent to which an evaluation can impact the teacher's methodology, assessment strategies, institutional policies, etc. *Authenticity* has to do with how much an evaluation resembles something that the student would do with the language outside of the classroom context (Brown & Abeywickrama, 2010; Combee et al., 2007; Rogier, 2014). Lastly, *transparency* means the degree to which students are informed about the expected goals and skills to be accomplished and the means through which these will be assessed (Rogier, 2014). But in addition to these principles, Brown & Abeywickrama believe that alternatives in assessment, such as the ones listed below, call on teachers to perform new instructional and assessment roles (2010, p. 123). In the authors words,[3] these forms of evaluation:

1. require students to perform, create, produce, or do something;
2. use real-world contexts or simulations;
3. are nonintrusive in that they extend the day-to-day classroom activities;
4. allow students to be assessed on what they normally do in class every day;
5. use tasks that represent meaningful instructional activities;
6. focus on processes as well as products;
7. tap into higher-level thinking and problem-solving skills;
8. provide information about both the strengths and weaknesses of students;
9. are multiculturally sensitive when properly administered;
10. ensure that people, not machines, do the scoring, using human judgment;
11. encourage open disclosure of standards and rating criteria; [and]
12. call on teachers to perform new instructional and assessment roles.

The Status Quo of Authentic Assessment

For the purpose of clarity, we will use the term authentic assessment throughout the whole paper, although we know that other terms such as on-

[3] Semicolons and period added to account for grammatical accuracy; original citation did not include any of these.

going assessment, alternative assessment, classroom assessment, qualitative assessment, or process assessment are also common in the professional literature. According to O'Malley & Valdez, alternative assessment means evaluation that is criterion-referenced rather than dictated or imposed by standardized measures. This type of assessment is authentic in that it resembles the types of activities used in the classroom and in real-life contexts; and it is also "consistent with classroom goals, curricula, and instruction" (O'Malley & Valdez, 1996, p. 2). Simply put, authentic assessment refers to "the multiple forms of assessment that reflect student learning, achievement, motivation, and attitudes on instructionally-relevant classroom activities" (O'Malley & Valdez, 1996, p. 4).

This new perspective emerges from a need to more consistently assess the full range of skills in students, as well as to parallel the advancements gained in recent decades in the field of language education. It also arises from the realization that the insights of traditional assessment contribute little to curriculum development (O'Malley & Valdez; Rogier, 2014). In modern days, such concerns have also been backed up by policy makers and administrators who have become aware that traditional assessment (particularly multiple-choice-based tests) does not help develop the higher order skills that students will need in order to meet the challenges that await in the decades to come (Brown, 2010; O'Malley & Valdez, 1996).

Authentic assessment includes a wide array of alternatives such as portfolios, student self-assessments, performance assessment, writing samples, projects and exhibitions, experiments or demonstrations (O'Malley & Valdez, 1996), rubric-referenced assessment, conferences and interviews, observations, peer-assessments, and journals such as language-learning logs, response to readings, dialogue journals, acculturation logs, etc. (Brown & Abeywickrama, 2010).

In order to design this type of evaluation, O'Malley & Valdez (1996, pp. 17-19) have proposed some crucial steps such as the following:

1. build a team;
2. determine the purposes of the authentic assessments;
3. specify objectives;
4. conduct professional development on authentic assessment;
5. collect examples of authentic assessments;
6. adapt existing assessments or develop new ones;
7. try out the assessments; [and]
8. review the assessments.

When used appropriately, authentic assessment can yield invaluable benefits, particularly in providing raw material for reflection and decision making. It also has great potential for transparency since authentic assessment demands that evaluation criteria be informed in advance to students. But above these gains, the noblest of all is perhaps that it helps bring effective teaching and assessment together. In our opinion, when educators manage to accomplish this, we have to agree with Tudor (2001) that the language teaching field has reasons to be proud of the advancements made in the last four or five decades.

Naturally, challenges exist which need to be acknowledged and dealt with. One of them is certainly the issue of maximizing practicality and washback (Brown, 2010). While authentic assessment guarantees high levels of washback, authenticity and appeal to intrinsic motivation, it tends to be largely impractical (time consuming and difficult to handle in some contexts) and unreliable (Brown, 2010). Added to that is the fact that students may not be ready for a shift in evaluation trends and may therefore experience distress in adjusting to them, not to mention the likely resistance from the administration—especially since authentic assessment demands a lot of willpower and resources. Additionally, these alternatives in evaluation tend to be highly subjective, especially if training is not made available to teachers and assessors. However, a question we need to ask here is, are traditional assessments free from this type of subjectivity? Grant (1990, p. 5) has provided the following insights on the matter:

> Though the scoring of standardized tests is not subject to significant error, the procedure by which items are chosen, and the manner in which norms or cut-scores are established is often quite subjective--and typically immune from public scrutiny and oversight.

In spite of the theoretical momentum gained so far by the field of authentic assessment, advancements are much less solid as we move on to the specifics of authentic listening assessment. In fact, things are not even clear-cut for the *teaching* of listening due to historical disagreements on what the construct itself should involve, a condition which has earned listening the nickname of "the Cinderella of communication strategies" (Vandergrift, 1997). While empirical studies have been and continue to be reported on authentic listening assessment (e.g., Ghaderpanahi, 2012; Pan, 2017; Porter & Roberts, 1981), findings as to its real benefits are scarce within the macro picture of language assessment and ELT.

After examining the complex —and often polarizing— status quo of authentic assessment, to a large degree we agree with Brown's (2010) assertion that we must "scrutinize the practicality, reliability, and validity of assessment alternatives while simultaneously celebrating their washback potential,

authenticity, and appeal to students' intrinsic motivation" (p. 126). This must be done, ideally, in the understanding that: (1) in time, authentic assessment needs not be so time-consuming and unreliable and (2) the long-run benefits of authentic assessment outweigh its potential drawbacks. In the current research project, we have made all possible efforts to integrate the cornerstones of assessment —usefulness, purpose, reliability, validity, practicality, washback, authenticity, transparency, and security (Brown, 2010; Combee et al., 2007; Rogier, 2014)— with Brown & Abeywickrama's (2010) principles for authentic assessment, as well as with O'Malley & Valdez's (1996) steps to develop this kind of assessment.

Methodology

This study can be characterized differently based on its paradigm, epistemology, and design. Following Cohen, Manion, & Morrison's view, our investigation adheres to a naturalistic paradigm since it conceives truth as personal, subjective and unique and thus deems research as depending on several perspectives, including the subjects', the researchers' and society's (2007). The inquiry is therefore based on an emic epistemology, where knowledge is co-constructed from within the culture, as opposed to outsider-expert research which seeks to create universal laws about the phenomenon under study (Ecksenberger, 2014). As naturalistic and emic, our study uses a QUAL-Quan research design, with a strong prevalence of qualitative data. Also known as the "exploratory mixed methods design," this model begins with a qualitative phase to get a first glimpse of the research problem and then weights results against quantitative data (Gay, Mills, & Airasian, 2009, p. 463), which, incidentally, accounts for methodological trustworthiness.

Participants and Context

Participants comprised thirty-eight purposively-sampled junior students enrolled in an English teaching major at UCR, West Branch. The program was first opened in the 1970s as a response to the growing demand for English teachers and has undergone adjustments in both curricular content and administrative matters ever since. In 2015, the major was accredited by Costa Rica's official accrediting entity (SINAES)[4] and is currently being evaluated to pursue reaccreditation in 2019. In the light of these processes, certain teaching changes were suggested, including authentic assessment as part of the evaluation criteria in all courses. This led the researchers to use authentic

[4] Spanish for *Sistema Nacional de Acreditación de la Educación Superior*.

assessment as a platform for the course *IO-5005 Laboratorio de Comunicación Oral V* (henceforth, Lab V), whose goal is to develop listening and vocabulary skills for academic purposes as a complement to the *IO-5440 Comunicación Oral V* course. Lab V covers a range of top-down and bottom-up skills such as listening for main ideas, making inferences and drawing conclusions, listening for details, discriminating between phonemes, and many others. In addition to this, it includes preparation exercises for the Test of English as a Foreign Language test (TOEFL) to train students in the test-taking skills they will need if they are required to take this examination later in their professional lives. Historically, the TOEFL component, in particular, has represented a significant challenge for English teaching learners, with verbal reports of anxiety, stress, and even academic frustration.

Research Procedures

In order to conform to the new accreditation requirements and account for a more democratic approach to language instruction, several techniques were combined in an on-going assessment project. These did not seek to replace summative evaluations, but simply to bring in a wider spectrum of assessment possibilities into the classroom. The data were collected at different points in the semester and later cross-checked for content validity. Participants were asked for written consent and guaranteed full confidentiality to ensure response accuracy and comply with research ethics. Ethics procedures included, amongst others: (1) assigning code names to participants so that their real identities were protected, (2) keeping data sources such as portfolios, student self-assessments and checklists in a locked file cabinet in the personal possession of the researchers, and (3) offering to share a copy of the final manuscript before it was sent for publication.

Data Collection Instruments

- **Plans of Improvement (POIs):** On a weekly basis, informants were asked to (1) set their own learning goal, (2) design an activity to reach that goal, (3) set a schedule to carry out the proposed activities, (4) assess the effectiveness of the activities, and (5) write annotations for the following week in order to achieve a sense of content continuity and write down notes that they wished to keep in mind for the following week. Voluntarily, students sent their POIs to get professor's feedback that helped build interactive-diaries dynamics (see Genesee & Upshur, 1996), which in turn fostered closer communication between the students and the instructor and allowed for classroom adaptations based on the POI reports.

- **Student Self-Assessments:** Three times during the semester, students filled out a self-assessment form to reflect upon their progress and weaknesses and devised strategies to overcome such deficiencies. They were advised to link this technique to the POIs and peer-assessments they did at other points during the term.
- **Verbal Calls:** Each time summative assessments were returned, students were encouraged to give impromptu reflections on test scores and ways to improve their listening skills. They were also asked to provide feedback on the quality of the assessments.
- **Final Portfolio Report:** As a final project, informants compiled all their POIs in a learning portfolio. They made corrections suggested along the course of the semester and wrote an introduction and a conclusion where they reflected critically about the experience in the project.
- **Strategy-Assessment Checklist:** At the end of the semester, participants rated the effectiveness of the project in terms of listening comprehension via a checklist that included statements on whether and to what extent the project had helped achieve a series of cognitive and metacognitive skills.
- **Teacher-Student Conferences:** These were five-minute meetings for students to verbalize their experiences with the authentic assessment strategy. These conferences served as an opportunity for students to discuss the benefits and challenges of the above techniques, and as a means for the professor to provide one-on-one advice on the strategies students were putting into practice.

Trustworthiness Methods

Three procedures were used to achieve instrument and methodological trustworthiness: (1) employing multiple data collection techniques (see previous section), (2) using various levels of triangulation (methodological, theoretical, and researcher triangulation), and (3) self-disclosure or reflexivity. The latter included being critical about how the researchers' own biases may influence the study; thus, hidden beliefs about authentic assessment were discussed and calibration sessions were conducted to agree upon the naming, grouping, and displaying of the categories.

Data Analysis and Interpretation

As stated earlier, this study sparked off as to assess the benefits and implications of bringing authentic assessment into listening comprehension courses at an EFL program in Costa Rica. As data were named, grouped, and

displayed for analysis, three major categories emerged in terms of benefits for students: *self-awareness and goal setting*, *sense of achievement*, *critical thinking*, and *general knowledge*. In the analysis ahead, qualitative data will be interpreted concurrently with quantitative information gathered via the participants' appraisal of the project. The following citing codes will be used to specify the instrument and sources where data come from.

Table 1. Qualitative Data Citation Codes

Instrument Type	Data Source	Citing Code
Plans of Improvement (POIs)	Oral Lab V Students	POI 01-036
Student Self-Assessment	Oral Lab V Students	SSA 01-036
Student Portfolios	Oral Lab V Students	SP 001-036
Strategy-Assessment Checklist	Oral Lab V Students	SAC 01-036

Source: Researchers' own design

For each category, qualitative excerpts of raw data will be accompanied with verbal descriptions and explanations. These will be further triangulated with the participants' quantitative appraisal of the project in general. Different data sources suggest the project had a positive impact on at least four major areas: self-awareness and goal-setting, sense of achievement, critical thinking, and general knowledge.

Self-Awareness and Goal-Setting

The first recurrent theme was self-awareness and goal-setting. Participants reported that keeping a weekly plan of improvement helped them become aware of both their limitations and accomplishments, which in turn led them to set goals to tackle learning problems or to stick to the tasks they found beneficial. In his accounts of the events for week three, participant 26 acknowledges some limitations he faced but immediately goes on to state that he will take advantage of the experience in order to maximize his learning on the following week:

> In short, I must manage to control the environment and myself to perform the subsequent activities in an accurate way. Besides, I will use all the techniques I have learned, but did not applied, to do better on the next test practice (POI 26 [Week 3], sic).

Participant 38 offers valuable evidence in this same regard throughout different POIs she created. On week two, she states that she chose to work on identifying main ideas and details in the topic of *brain and language* because she had previously identified limitations on these two skills: "In the second video I missed some details, and I could not understood very well some aspects that he was talking about so I played it twice" (POI 38 [Week 2], sic). On the following week, she issues the following reflection: "This week I decided to practice TOEFL listening exercises because I noticed that I could identify general ideas when I practice with videos at home, but it was difficult during the practice at the lab" (POI 38 [Week 3], sic). Shortly after, she explains that the difficulties experienced in class were due to concentration problems resulting from lack of vocabulary:

> It continues being difficult for me to be concentrate while listening to long audios. Moreover, I notice that my lack of vocabulary sometimes affect me while listening an audio. Because when I do not understand a word, I miss my concentration by monitoring what the word means or thinking if I know the word (POI 38 [Week 3], sic).

Just like participant 26, informant 38 also engages in goal-setting upon becoming conscious of existing limitations: "This week I listened to lectures and conversations from TOEFL exercises in order to improve concentration, fast answering and details" (POI 38 [Week 4]). If we notice, on week two this participant becomes aware of her need to work on listening for details, on week three she identifies weaknesses in her concentration, and on week four she tackles the two issues together and goes beyond by working on TOEFL exercises. The reliability of this data is further corroborated via the participant's thoughts in the final portfolio:

> …as TOEFL is a type of exercise which is really fast, and I had problems with the speed of the audios, I discovered that writing down what I was hearing was a technique to improve my concentration and my speed to answer that help me a lot (SP- 038, sic).

Evidence of self-awareness also comes from participant 22, who reports to have realized the wealth of free access listening tools available online:

> Thanks to the plan of improvement, I was able to find helpful activities, applications and websites that are really helpful to develop better skills […] I now realized that there are excellent tools on the internet […] and are free to use, so there is no excuse not to go further in my learning process (SP- 022, sic).

A common denominator within the current category is the participants' awareness in relation to the TOEFL component. This came up most evidently in the self-assessment instruments they completed, with comments such as "the major challenge I'm facing right now is the TOEFL exercises […]" (SSA 33, sic); "in those exercises the time is reduced so I don't have enough time to read the answers before completing" (SSA 24, sic); or, "the major challenge is TOEFL exercises […] There are too fast and also idioms and phrases are challenging" (SSA 34, sic). In all cases, however, the constant among respondents was also that they had goals underway in order to meet these drawbacks. Comments on this included, for example, "I'm looking for audios similar to TOEFL test" (SSA 31, sic); "to overcome these challenges I need to practice TOEFL at home" (SSA 33, sic); "I became a better listener […] I just need to devote more time on the challenges" (SSA 34, sic); and "[I will] spare some time for practicing TOEFL and work on the portfolio" (SSA 37, sic).

In addition to the above, a variety of claims were made that sometimes the activities did not work as expected. Participant 33 describes limitations in terms of the content and length of a video that she meant to watch in order to reinforce TOEFL test taking skills. As she puts it:

> This activity was too bored to me because the passage was too long about 6 minutes and I get lost in the middle. Also, I do not like that the video present the questions but not the answer of each question so I do not know if I am answering correctly (POI 33 [Week 7], sic).

Something that merits our attention is the honesty element. Acknowledging these facts is relevant to the validity of our results as they evidence participants' confidence to write about both, the cases where activities did not work, and about their thoughts on what teachers should consider for further project implementation. At any rate, the mere act of verbalizing these issues is but proof that self-awareness has indeed taken place.

Turning now to the quantitative data on this category, results are generally congruent with the ones drawn from the qualitative information. On assessing the correlation between the project and the development of self-awareness and goal-setting to tackle various learning limitations, 76.3% ranked it as "very much" (beneficial), 21.1% rated it as "about right", and 2.6% evaluated that link as being "too little"; none of the informants ranked it as totally useless. A chart with these numerical values will be presented at the end of this data analysis.

Sense of Achievement

The second category has to do with student's reported sense of achievement. From improved tactics for the TOEFL test to the mastery of regular course content, data suggests clear progress in terms of this category:

> […] I got into a higher difficulty of audios, and I learned to have a discipline to practice it without problem… at the beginning of the course I got a really low grade in my diagnostic test, and since that moment I felt afraid and nervous about TOEFL exercises. That is why, I think I took around too weeks to start to practice deeper with TOEFL audios, but now I can say that I faced my fears about that, and at this moment I feel comfortable practicing with that kind of audios (SP- 038, sic).

On the same lines, participant 21 states that despite some difficulties he faced, he made substantial progress because of the project:

> In spite of some difficulties; through these ten weeks I could see significant progress in my listening skills due to this plan. I believe that the portfolio strategy really works; it is a useful tool because it allows a detailed sight of weaknesses, and main points to work on. I am going to apply this strategy from now on in order to get a better understanding of English in all areas of my life (SP- 021, sic).

Participant 06 takes it one step beyond and analyzes the possibility of replicating the POI strategy in other classes in order to test its effectiveness in further educational contexts. In her final portfolio report, she notes:

> I think that this type of assessment is very enriching for the teacher and for the student, it is a way to document achievements, where the student describes his way of learning, expresses doubts and make comments on each activity, all this makes the portfolio an interesting and enriching tool. I have the intention to apply this methodology in other subjects and verify their effectiveness and efficiency in any educational program (SP-006, sic).

Further evidence on sense of achievement —especially in terms of vocabulary acquisition and listening strategies— was recorded through the student self-assessment strategy, with the inclusion of comments such as "I learned more vocabulary, phrasal verbs. I also learned to listen for main ideas and details […]" (SSA 15, sic); "I learned vocabulary and a strategy that is helping me with TOEFL exercises" (SSA 18); and "I could comprehend easily

details in audios, and I could use new vocabulary" (SSA 01, sic). From the final portfolio, similar general comments were documented:

> Without more to say, I really thank God and the professor for having this wonderful idea of the weekly plan. It helped me a lot because even when I like English, I know that I wouldn't have been as constant practicing it as I was with this work (SP- 034, sic).

In addition to these two sub-components, sense of accomplishment was emphasized also in terms of skills integration, for the different assessment strategies required students to combine a number of linguistic skills. In participant 19's words, "I have learned more vocabulary and finally, I have practiced my writing skills every time I had to write the evaluation of the activities" (SP- 019). To sum up, we end this sub-section with a comment from subject 20, who outlines the main advantages of the project:

> Sincerely, I am motivated to do those exercises because I really have seen progress during my last two evaluations. I feel like "super powerful" and a feeling of support have come to me. It really says "go Luis, you can do it". And, indeed, I am doing it. […] Finally, I would like state that during my three years being tested in my laboratory courses, this listening assessment plan is one on the most brilliant ideas to put into practice. All the changes that I have experimented during this plan have been successful (SP-020, sic).

On the quantitative side of the data, numbers from the strategy-assessment checklist (SAC) reinforce the qualitative analysis above. Thirty-six-point eight percent of the informants evaluated the project as "very much" helpful in building sense of achievement, 60.5% as "about right", 2.6% as "too little", and 0% as not useful at all (see full chart at the end of data analysis section). Although not conclusive, results from this category match previous findings by Sevilla and Gamboa (2016a) that this type of pedagogical innovations fosters student sense of achievement and bring them closer to the life-long learning expected at various levels of their professional profile.

Critical Thinking

The third category was critical thinking; and while we are fully aware that critical thinking is a rather elusive construct, for the purpose of this analysis we understand it simply as the ability to use higher order skills including assessing, valuing, appraising, criticizing, weighing, and recommending, as suggested in various critical-thinking-relevant phases of Bloom's taxonomy

(see Bers, 2005). Thus, our analysis of critical thinking begins with evidence from participant 29, who praises the strategy as a whole but criticizes the course syllabus, the students, and herself:

> I did not have many complaints about the course nor the professors but I complain on the program and the students including myself. I consider it is necessary to change the material for Lab courses because unless we learn with prerecorded conversations natural ones will it give us the real sense of communication […] In my own opinion, students should be doing this extra effort not as a part of the evaluation but for personal improvement (SP-029, sic).

Strictly speaking, besides critical thinking this excerpt also shows proof of self-awareness. However, evidence on critical thinking predominates as the student is critical of the course material, the syllabus, herself and her classmates. She also weighs and recommends changes on ETM's listening courses in order to make learning more authentic. In a similar context, participant 33 is critical of her own choice of materials for TOEFL practice when she reports: "The activity of watching this video supposed to help me practicing TOEFL exercises […] but this activity do not really work to me because that video was too long and does not have the answer of each question […]" (POI 33 [Week 7], sic). Another participant (36) acknowledges that the reason she was successful in completing the project was that it had a summative value. This is so —she admits— due to the academic workload it places on students: "[…] If the portfolio had not had a score I think I would not have done it because it is too difficult to spend time doing these activities because I had many things all the weeks" (SP-036, sic). And still another participant takes the lead and provides recommendations for future implementation:

> I recommend both professors and students that if they could share information about websites or educational applications in order to do the plan that would be great, because sometimes students have a really hard time finding acceptable activities and they might not know how to search for that on the Internet (SP-022, sic).

A last comment comes as a word of motivation for students interested in a project of this kind in the future:

> I would suggest not giving up when the exercises become tougher or when the person is not so motivated; quite the contrary, people must be aware that there are good and bad days when they may either do better than they have done before or mess all up (SP-25, sic).

At a numerical level, data displays high degrees of consistency with the qualitative analysis. Students' rating of critical thinking skills development was 60.5% for "very much", 31.6% for "about right", 5.3% for "too little", and 2.6% for "nothing". Of the three categories analyzed so far, critical thinking exhibits one of the highest "very much" rankings; however, it is also the first to exhibit any values for "nothing". The instruments used did not seek to delve into the reasons for these ratings, but data from the strategy assessment checklist indicates that critical thinking was particularly promoted in the TOEFL exercises because many exercises dealt with implied meaning and inferences, and that whenever it was not promoted, it was because some of the activities did not exhibit this component explicitly.

General Knowledge

This is the last category of analysis identified in terms of benefits of authentic listening assessment, with qualitative data suggesting promising outcomes at various levels. Topics such as brain functioning, U.S. politics, cultural stereotypes and physical disabilities, interiors decoration, and insights about law and teaching make up some of the achievements reported through different data collection instruments. On week two, participant 38 offers a claim along these lines as she states: "Moreover, the content of the video helped me to understand better how my brain works, and how it relates to my language learning process." (POI-038, sic). Along the same lines, informant 31 provides similar evidence when, after outlining her main challenges with the plan of improvement, she asserts:

> However, not all was lost because I have had noticeable improvement in listening, for example I was able to understand some controversial topics around Donald Trump and his argumentations about *Latinos.* Also, when I watched the *My Name is Khan* I felt overwhelmed for all the Muslim polemic, and situations that people with special necessities must face (in the case of Khan). Additionally, I learned the reasons why emotions are produced and the psychology of dreams by watching *Inside Out* (POI-031, sic).

A fact that stands out from the data above is the variety of topics this student reports to have explored on a single week. Although we see a common thread between the first two topics (i.e., politics and ethnic groups), a different subject is reported on at the end of the paragraph (the origins of emotions), which once again points to the project's potential to help students devise their own learning agendas and examine the issues of their own preference. An additional example of this can be found in participant 37, who points out: "I

watched videos related to things which I like because when I was tired of academic life I spent time looking for something interesting and I found videos about how to decorate rooms and clothes," as well as from informant 20, who highlights:

> To be honest, this week has been the most effective because I have been looking for information that is not presented in our feel of teaching. So, I have learned different laws and statistics that different countries apply in their working context. I am interested in this topic because becoming a lawyer is one of my future goals. Also, I discovered that I learn more when something calls my attention. Even if I am not able to understand every single word I am able to understand what the laws establish (POI-020, sic).

This passage presents a blend of at least two elements: general knowledge and goal-setting. Although goal-setting was analyzed previously in this section, we point it out here as an appendage of general knowledge since, as we have interpreted it, the student's immediate motivation to explore these topics is part of a learning need experienced in the course. The convergence of other goals and topics is, at any rate, an added bonus in the context of any educational endeavor, and one which in our opinion should attract more attention from instructors and stakeholders alike.

As a last point, analysis of this category concurs with the results for the *sense of achievement* category above, in that they continue to parallel Sevilla and Gamboa's 2016a claim that teaching strategies of this kind serve as a platform to prompt not only the target contents but also learning autonomy.

Numerically speaking, results appear to be encouraging as well. On table 2 below, we sketch a summary of the percentages coming from the participants' quantitative appraisal of the project, and which we have used throughout the previous paragraphs to triangulate the information gathered. The numbers on the top indicate the extent to which informants felt the project had been beneficial in each of the categories discussed previously (1= nothing; 2= too little; 3= about right; 4= very much; NA: No Answer).

Table 2. Summary of Students' Assessment of the Project

Criteria	1	2	3	4	NA
The project has helped me to…					
1. become aware of my limitations and set goals to meet these limitations	0%	2.6%	21.1%	76.3%	
2. build a sense of achievement on contents proposed for the week	0%	2.6%	60.5%	36.8%	
3. reflect on areas where I have made progress	0%	2.6%	31.6%	65.8%	
4. develop a general sense of critical thinking	2.6%	5.3%	31.6%	60.5%	
5. acquire general knowledge on the topics studied	0%	5.3%	28.9%	65.8%	
6. organize my time and deal with homework assignments	2.6%	18.4%	42.1%	36.8%	
7. process the contents of the week	2.6%	13.2%	39.5%	44.7%	
8. improve my (listening) learning strategies	0%	0%	21.1%	78.9%	
In general, this experience…					
1. motivated me to undertake independent, life-long learning	0%	2.6%	15.8%	78.9%	2.6%
2. helped me to nurture permanent (listening) learning habits	0%	2.6%	21.1%	73.7%	2.6%

Source: Participants' responses

Despite these positive results, we are certain that they do not escape criticism. Skeptics could argue that such benefits are the natural product of students' immersion in the course; that the opposite would be rather counterintuitive given the amount of effort invested from beginning to end. Following this logic, one would need to admit that there is little or no credit in the outcomes reported herein. Nonetheless, while such arguments may be true in principle, the current study does not pursue credits of any sort but instead seeks to systematize the benefits from this intervention strategy in order to draw implications for theory and future classroom application. Together, the results of this section indicate that authentic assessment bears a number of advantages to the teaching of listening comprehension in EFL programs. The next section, therefore, moves on to discuss the findings in light of the research goal and presents the conclusions and implications for theory and practice.

Conclusions

As stated previously, the present paper systematizes the benefits and implications of bringing authentic assessment into the listening comprehension classroom. In terms of the benefits, findings suggest positive links between authentic assessment and these classes in terms of the following themes: self-awareness and goal-setting, sense of achievement, critical thinking, and general knowledge. They also indicate that projects like this not only bring assessment and instruction together, but they also provide opportunities for skills integration through teacher conferences, peer-assessments, verbal calls, and the many other techniques employed. Along with this, students are given opportunities to build learner autonomy as they take charge of their own responsibilities in their own time and at their own will. In the long run, this will lead to what Tudor (2001) has termed the political (or social) scope of autonomy: a kind of learning that transcends the classroom boundaries and ensures successful learning for life.

Taken together, the evidence from this research provides insights for both classroom application and the field of knowledge at large. For classroom use, these data invite a move from traditional testing to ongoing assessment, and from norm-referenced towards more criterion-referenced evaluation. While generalizations are not possible given the small sample used, at any rate the aim should never be a radical "no" to objectively-scored examinations (if changes are to be enforced), but instead a more balanced application, and in combination with other techniques, in order to match the many realities that converge in an ELT context. For the advancement of second language assessment (SLA), our study sheds light on three key issues. First, it expands the body of available empirical literature on authentic assessment in listening comprehension—a skill already acknowledged as neglected even at the teaching level (see Gamboa & Sevilla, 2013 and Vandergrift, 1997). Second, it examines the benefits and implications of incorporating authentic assessment in the listening comprehension classroom considering Combee et al. (2007) and Rogier's (2014) cornerstones of language assessment, Brown & Abeywickrama's (2010) principles for authentic assessment, and O'Malley & Valdez's (1996) steps for developing this kind of evaluation. Lastly, it sets the grounds for future inquiries and proposes research methods that may help reach a more comprehensive examination of SLA. In so doing, such inquiries should not discard challenges such as *The Hawthorne Effect*, a phenomenon where participants modify their behaviors as a result of being part of research (Porte, 2010, p. 103), and increasing the sample size to attain generalizability of results.

All in all, while the project has proven successful in the scope of this research, we must not ignore a series of drawbacks that need to be considered if successful implementation is to be attained. For instance, we know from

instructional experience and educational research that this type of evaluation is time consuming. It, therefore, requires a lot of willpower, support from authorities, and a constant adjusting and readjusting of our professional praxes. It also bears high degrees of subjectivity, which demands that teachers combine assessment strategies to cross-check student progress and counterbalance such shortcomings. Lastly, an eventual change of this sort brings about receptivity issues (i.e., the degree to which a teaching change is welcome or not) which require effective coordination between curriculum planners and stakeholders on the one hand, and between instructors and department heads on the other. These and other drawbacks should be considered in future investigations to offer a more comprehensive picture of the benefits and implications, as well as challenges of authentic assessment in listening comprehension courses.

In terms of research application, two major limitations were identified. The first one is the small sample used, which limits transferability of findings; the second one is the scope of the study: Since the inquiry was limited to benefits and implications of authentic assessment in a listening comprehension class, it was not possible to assess the project's downsides for students, professors, policy makers, and curricular authorities. Future studies must definitely address these weaknesses and provide sound evidence on their outcomes.

Within this context, agnostics may wonder why we should adopt assessment models that cannot deal with these limitations successfully. Without meaning to oversimplify matters, perhaps the questions that need to be asked here are: Why continue to endorse traditional models that have not only proven insufficient but also ignored the wealth of other alternatives and failed to deal with these same issues effectively? What if instead of arguing over the traditional-versus-authentic-assessment dichotomy we joined efforts to make language assessment a more democratic, more participatory, and more honest endeavor? For now, these are questions that will only be answered through systematic, empirical research on the links between authentic assessment and listening comprehension in EFL programs.

References

Airasian, P. W. (2001). *Classroom Assessment: Concepts and Applications*. Boston, MA: McGraw Hill.

Ali, I., & Ajmi, A. (2013). Towards Quality Assessment in an EFL Programme. *English Language Teaching*, *6*(10), 132-148.

Aliweh, A. M. (2011). The Effect of Electronic Portfolios on Promoting Egyptian EFL College Students' Writing Competence and Autonomy. *The Asian EFL Journal, 13*(2), 90-132.

Bailey, K. M., Curtis, A., & Nunan, D. (2001). *Pursuing Professional Development: The Self as a Source.* Boston, MA: Heinle & Heinle.

Bers, T. (September 06, 2005). Assessing Critical Thinking in Community Colleges. *New Directions for Community Colleges, 2005, 130*, 15-25.

Brown, H. D., & Abeywickrama, P. (2010). *Language Assessment: Principles and Classroom Practices.* White Plains, NY: Pearson Education.

Charvade, M. R., Jahandar, S., & Khodabandehlou, M. (2012). The Impact of Portfolio Assessment on EFL Learners' Reading Comprehension Ability. *English Language Teaching*, *5*(7), 129-139.

Cohen, A. D. (1994). *Assessing Language Ability in the Classroom*. Boston, MA: Heinle & Heinle.

Cohen, L., Manion, L. & Morrison K. (2007). *Research Methods in Education* (6th Edition). New York: Routledge.

Creswell, J. W. (2007). *Qualitative Inquiry and Research Design: Choosing Among Five Approaches* (2nd edition). Sage Publications, Inc.

Dewi, R. C. (2018). Utilizing Authentic Materials on Students' Listening Comprehension: Does it have Any Influence? *Advancements in Language and Literary Studies*, *9*(1), 70-74.

Diller, K. R., & Phelps, S. F. (2008). Learning Outcomes, Portfolios, and Rubrics, Oh My! Authentic Assessment of an Information Literacy Program. *Portal: Libraries and the Academy*, *8*(1), 75-89.

Douglas, D. (2000). *Assessing Language for Specific Purposes*. Cambridge, UK: Cambridge University Press.

Frey, B. B., Allen, J. P., & Schmitt, V. L. (2012). Defining Authentic Classroom Assessment. *Practical Assessment, Research and Evaluation, 17*(2), 1-18.

Gamboa, R., & Sevilla, H. (2013). Assessment of Listening Comprehension in Public High Schools of Costa Rica: The West and Central Pacific Case. *Proceedings of the 11th Hawaii International Conference on Education, Honolulu, Hawaii.* 1-24.

Gamboa, R., & Sevilla, H. (2016a). The Impact of Teacher Training on the Assessment of Listening Skills. *Revista LETRAS*, *57*, 77-102.

Gay, L.R., Mills, G.E., & Airasian, P. (2009). *Educational Research: Competencies for Analysis and Applications*. New Jersey: Pearson Education, Inc.

Ghaderpanahi, L. (2012). Using Authentic Aural Materials to Develop Listening Comprehension in the EFL Classroom. *English Language Teaching, 5*(6). 146-153. doi:10.5539/elt.v5n6p146

Genessee, F., & Upshur, J. (1996). *Classroom-Based Evaluation in Second Language Education*. Cambridge: Cambridge Language Education.

Grant, W., & Educational Resources Information Center (U.S.). (1990). *The Case for Authentic Assessment.* Washington, DC: U.S. Dept. of Educational Research and improvement, Educational Resources Information Center.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A Practical Guide to Alternative Assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Li, C. H. (2013). They Made it! Enhancing University-Level L2 Learners' Listening Comprehension of Authentic Multimedia Materials with Advance Organizers. *The Asia-Pacific Education Researcher, 22*(2), 193-200.

McKay, P. (2006). *Assessing Young Language Learners*. Cambridge, UK: Cambridge University Press.

Miller, L. (2003). Developing Listening Skills with Authentic Materials. *ESL Magazine*, *6*(2), 16-18.

Moya, S. S., & O'Malley, J. M. (1994). A Portfolio Assessment Model for ESL. *The Journal of Educational Issues of Language Minority Students*, *13*, 13-36.

Murphy, V., Fox, J., Freeman, S., & Hughes, N. (2017). "Keeping it Real": A Review of the Benefits, Challenges, and Steps Towards Implementing Authentic Assessment. *All Ireland Journal of Teaching and Learning in Higher Education (AISHE-J), 9*(3). Retrieved from http://ojs.aishe.org/index.php/aishe-j/article/view/280

O'Malley, J. M., & Valdez, L. (1996). *Authentic Assessment for English Language Learners: Practical Approaches for Teachers*. Reading, Massachusetts: Addison-Wesley Publishing Company.

Pan, Zhixin. (2017). Assessing Listening for Chinese English Learners: Developing a Communicative Listening Comprehension Test Suite for CET. London and New York: Routledge.

Porte, Graeme. (2010). *Appraising Research in Second Language Learning: A Practical Approach to Critical Analysis of Quantitative Research.* Philadelphia: John Benmamins.

Porter, D., & Roberts, J. (January 01, 1981). Authentic Listening Activities. *ELT Journal (United States), 36*(1), 37-47.

Rogier, D. (2014). Assessment Literacy: Building a Base for Better Teaching and Learning. *English Teaching FORUM*, *52*(3), 2-13.

Rojas, O. (2004). El *portafolio* y la evaluación del proceso en traducción. *Revista LETRAS*, 36, 27-64.

Sevilla, H., & Gamboa, R. (2016b). Student Self-Evaluation and Autonomy Development in EFL Learning. *Revista de Lenguas Modernas*, *25*, 199-222.

Tudor, I. (2001). *The Dynamics of the Language Classroom*. Cambridge: Cambridge University Press.

Vandergrift, L. (1997). The Cinderella of Communication Strategies: Reception Strategies in Interactive Listening. *The Modern Language Journal, 81*(4), 494.

Yurdabakan, I., & Erdogan, T. (2009). The Effects of Portfolio Assessment on Reading, Listening and Writing Skills of Secondary School Prep Class Students. *Journal of International Social Research*, *2*(9), 526-538.

Authors

***Henry Sevilla Morales** holds an M.A. in Second Languages and Cultures. He has been an EFL instructor for 13 years and currently works at *Universidad Nacional*, Costa Rica. His research has been presented in numerous national and international conferences and published in various journals and conference proceedings in Costa Rica and The United States. His current research areas include reflective writing, learner autonomy, authentic assessment, critical applied linguistics, and translation studies.

ORCID: https://orcid.org/0000-0003-4040-8062

Lindsay Chaves Fernández holds an M.A. in Second Languages and Cultures from *Universidad Nacional,* Costa Rica. She has been an EFL instructor for over fifteen years and currently works at *Escuela de Literatura y Ciencias del Lenguaje*, *Universidad Nacional*. Her papers appear in different scientific journals and her research has been presented in national and international conferences in Costa Rica, Cuba, Canada, Spain and Switzerland. Among her research interests are composition and rhetoric, Information and Communication Technologies (ICT) in teaching, learners' diversity and motivation, translation studies, approaches to EFL teaching, and accreditation practices in higher education.

ORCID: https://orcid.org/0000-0001-7936-8112

How to reference this article: Sevilla Morales, H., & Chaves-Fernandez, L. (2019). Authentic Assessment in the Listening Comprehension Classroom: Benefits and Implications. *GIST – Education and Learning Research Journal*, *19*, 6-30. https://doi.org/10.26817/16925777.704