The Impact of Teacher Training on the Assessment of Listening Skills[1]

(El efecto de la capacitación en la evaluación de competencias auditivas)

Roy Gamboa Mena[2] Universidad de Costa Rica *Henry Sevilla Morales*[3]
Universidad Nacional, Universidad de Costa Rica

Abstract

The correlation between teacher training and the listening assessment practices is analyzed with regard to teachers in the Costa Rican educational system. The results of a workshop on theory of listening assessment and guidelines provided by the Ministry of Public Education suggest that these testing practices can be improved with training that unifies official guidelines and listening assessment theory.

Resumen

En el estudio se analiza la correlación entre capacitación docente y las prácticas de evaluación auditiva en el profesorado del sistema educativo costarricense. A partir de un taller sobre teorías de la evaluación auditiva y directrices al respecto, del Ministerio de Educación Pública, se colige que esas prácticas deben mejorarse mediante capacitaciones que unifiquen tales directrices oficiales y la teoría de esta área del saber.

1 Recibido: 10 de diciembre de 2014; aceptado: 5 de febrero de 2015. Elaborado a partir de una ponencia presentada en el I Congreso Internacional de Lingüística Aplicada (ConLA-UNA), llevado a cabo entre el 4, 5 y 6 de febrero 2013 (Brunca Extension, UNA).

2 Escuela de Lenguas Modernas (ELM). Correo electrónico: roy.gamboamena@ucr.ac.cr

3 Escuela de Literatura y Ciencias del Lenguage y ELM. Correo electrónico: henry.sevillamorales@ucr.ac.cr

Keywords: assessment methodology, language assessment, evaluating listening skills
Palabras clave: metodología de la evaluación, evaluación lingüística, evaluación auditiva

Introduction

Over the past decades, the assessment of listening skills in ESL and EFL learning has become an area of concern in both research and teaching. Such concern stems, by and large, from the fact that listening assessment has proven to be both a difficult area of language teaching and a relatively neglected field worldwide.[4] As for the context of the Costa Rican public education system, the state university board (*Consejo Nacional de Rectores*, CONARE) and four State universities (i.e., Universidad de Costa Rica, Universidad Estatal a Distancia, Instituto Tecnológico de Costa Rica, and Universidad Nacional) have provided training on language assessment for MEP in-service teachers ranked C1 according to the Common European Framework of Reference for Languages (CEF).[5] Because that training is offered only to teachers ranked C1, a large number of educators cannot receive it, and yet they are carrying out listening assessment in their institutions. That is, this group of teachers needs to assess listening without having been given any guidelines by the Ministry of Public Education (MEP).[6]

Thus, the present study is based on the need to fill the gap between listening assessment as a neglected language area, the MEP's overt lack of specific listening assessmet guidelines, and the actual

4 D. J. Mendelson, *Learning to Listen: A Strategy-Based Approach for the Second Language Learner* (San Diego: Dominie Press, 1994); L. Vandergrift, "The Cinderella of Communication Strategies: Reception Strategies in Interactive Listening," *The Modern Language Journal* 81 (1997): 494-505; N. Osada, "Listening Comprehension Research: A Brief Review of the Past Thirty Years, *Dialogue* 3 (2004): 53-66; R. Gamboa and H. Sevilla, "Assessment of Listening Comprehension in Public High Schools: The West and Central Pacific Case," *Proceedings of the 11th Hawaii International Conference on Education*, 6-11 Jan. 2013, Honolulu, Hawaii, 2013.

5 Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, and Assessment* (Cambridge: Cambridge University Press, 2001).

6 Gamboa and Sevilla.

test-design practices of in-service teachers. In addition, it stems from the need for further teacher training on the effective application of assessment principles.

The results of this study indicate that listening assessment practices of EFL teachers can be improved significantly if training that combines assessment theory and institutional guidelines on assessment is provided. Likewise, these results suggest that further training is needed so that all existing gaps between theory and practice can finally be closed up. Within the field of Applied Linguistics, the study expands the existing body of literature on the subject of listening assessment, and it opens an avenue for further research in this area so that MEP educational policies can be oriented towards consistency between theory and practice. In an age of globalization in which English skills are paramount for effective multicultural communication, and in which active listening skills have become vital in language learning, research on listening assessment proves not only relevant but also crucial as a way to provide insights on how to improve teaching in the context of English as a Foreign Language.

Review of the Literature

A Brief History of Listening Assessment

The history of listening assessment can be traced back to the development of two currents of language teaching approaches that appeared during the second half of the twentieth century; that is, the audiolingual method and the communicative approaches to language teaching. Thus, Buck[7] described three approaches to listening assessment. The first was developed during the 1950s as the audiolingual method came into existence. In this approach, according to Coombe et al.,[8] listening was broken down into separate elements to be assessed.

7 G. Buck, *Assessing Listening* (Cambridge: Cambridge University Press, 2001).

8 C. Coombe, K. Folse and N. Hubley, A *Practical Guide to Assessing English Language Learners* (Ann Arbor, MI: University of Michigan Press, 2007).

The rationale comes from the belief that "it was important to be able to isolate one element of language from a continuous stream of speech and that spoken language was believed to be the same as written language […]."[9] The second approach, called the *integrative approach*, came into existence in the 1970s. Coombe et al. argue that tests in this approach sought to assess the learner's capacity to use many language bits at the same time. The whole of a language was seen as being greater that the sum of its parts. The last approach to listening assessment, as proposed by Buck, is to be found within the communicative approach to language teaching developed during the 1970s when "the status of listening comprehension began to change from being incidental and peripheral to a status of central importance."[10] According to this approach, "the listener must be able to comprehend the message and then use it in context."[11] Nonetheless, these changes in listening assessment in the past decades have led to serious dilemmas for teachers and researchers, and evidence exists that listening itself has been neglected in many English programs.[12]

In the case of Costa Rica's public education system, the issue is especially challenging because teachers often find themselves facing discrepancies between what the theory says they should do in terms of assessment and what the MEP requests. These contradictions are found, for the most part, when it comes to task design. For instance, in theory[13] when designing multiple-choice tasks to assess listening, it is suggested that two distractors and one correct answer be included, whereas the MEP requests three distractors and one correct answer. The scenario is further complicated because the MEP dictates guidelines for general assessment, not for listening[14]; and yet English teachers

9 Coombe, Folse and Hubley, 9.
10 Osada, 5.
11 Coombe, Folse and Hubley, 91.
12 Mendelson; Osada; and Vandergrift.
13 For example, Coombe, Folse and Hubley.
14 See Ministerio de Educación Pública, *La prueba escrita* (San José: Departamento de Evaluación, 2011).

need to comply with them as if assessing language entailed the same procedures as assessing other subjects.

The unavailability of guidelines for listening assessment in Costa Rican schools poses another challenge for teachers. Gamboa and Sevilla claim that "so far, there are parameters for designing written tests in general, but there are no such parameters specifically for designing listening tests."[15] The implies that instructors need to follow directions based on a word-of-mouth tradition passed down from school to school and from one regional branch to the other. This results in teachers creating tests in different ways, thus undermining language learning and teaching.

The Role of Validity, Reliability, and Washback in Language Assessment

Recent literature on language assessment has devised eight basic cornerstones of language assessment (i.e., usefulness, validity, reliability, practicality, washback, authenticity, transparency, and security). However, in recent years, there has been an increasing interest in the importance of reliability, validity, and beneficial backwash, briefly described below.

Reliability refers to how reliable a test is; that is, how consistent the scores of a test are over time, or its ability to obtain the same or at least a similar score from the same student if the test is given by a different tester and at a different time. When it comes to test reliability, two key variables need to be distinguished: *intra-rater reliability* and *inter-rater reliability*.[16] Because the nature of language testing suggests the need to minimize rater's subjectivity while scoring, it is important to understand the purpose of these two variables. Intra-rater reliability, on the one hand, refers to a testing practice where only one rater is involved in the scoring process; usually, the teacher. Inter-rater reliability, on the other, involves more than one rater. The latter is

15 Gamboa and Sevilla, 3.

16 J. D. Brown, *Testing in Language Programs: A Comprehensive Guide to English Language Assessment* (New York, NY: McGraw Hill, 2005).

particularly difficult to achieve since the worldviews of many raters come into play, and lead to disagreements about what score to give on a test. Brown believes that one of the best ways to solve the issue is by using detailed and well-defined rating criteria and to have several training sessions where the examiners learn to apply the rating criteria as objectively and accurately as possible. Alderson,[17] however, warns that this may be achieved only if the teaching setting allows for such action and there is enough determination to achieve it.

Another element that has been given attention within the scope of language assessment is that of validity. Test validity deals with the extent to which a test measures what it purports to measure.[18] The test validity principle is comprised of several layers that need to be considered during assessment. The first such layer is *content validity*, which is about whether the test includes a representative sample of the content that it purports to measure.[19] This principle is often violated by many testers and curricular authorities as well, often without considering the implications behind it. One typical consequence is that students score low on the tests because they are assessed by means of a different skill than the one studied in class. Similarly, cases exist where the teacher uses methods in the class that do not match the ones used on the test. Another layer is *criterion-related validity*, in which the test designer compares his/her test with a well-known test, like the TOEFL or the TWE, and sees if the one s/he designed is close to it in quality. This can be useful also to confirm that his/her test is measuring the same constructs as the test used as reference (e.g., listening, writing, speaking, etc.). *Concurrent validity* is another important construct in testing validity. It follows the same principle of criterion-related validity, but administering both the test created by the teacher and the test used as a reference at the same time. Finally, *predictive validity* has to do

17 J. Alderson and L. Bachman, "Series Editors' Preface." In G. Buck, *Assessing Listening* (Cambridge: Cambridge University Press, 2001).

18 Coombe, Folse and Hubley.

19 Brown.

with the predictability value that a test may have in determining the test taker's success in a given communicative scenario. For instance, it is assumed that if a student scores high on the TOEFL, s/he will do well at an American university. In other words, a test must serve as a predictor of the students' success. All of these types of validity must be taken into consideration while interpreting test scores.

There are, however, some problems when dealing with validity. Probably the most typical one comes into play when testers or curricular authorities decide to administer standardized tests such as the TOEFL, the TOEIC, or the TWE for testing the skills of a particular group of students. Brown advises that if the tester wants to use a standardized test for a certain course, under particular circumstances, and for a specific group of students, adaptations need to be made so that validity principles are not violated.[20] As occurs with reliability, this requires a great deal of effort and determination, which might not always be a condition found in most public schools in Costa Rica.

As for beneficial backwash, this component has been given significant acknowledgement in the past years within the field of language testing. In general, it is seen as "the impact a test may have on learners and teachers, on educational systems, and on society at large."[21] It is gaining popularity among researchers and curricular authorities because, as discussed in literature on testing research, test scores may influence the decision-making procedures of a particular institution. Brown provides a practical and useful example of this phenomenon:

> Consider the following scenario: you are working in an institution that gets more funding if the number of students reaching a certain benchmark (i.e., standard) on the standardized test at the end of the year increases. As a result, at the end of the year, your director will be keeping tabs on how many of your students make the benchmark for funding. Do you think that will affect your teaching?[22]

20 Brown.

21 A. Hughes, *Testing for Language Teaching* (Cambridge: Cambridge University Press, 2003), 53.

22 Brown, 242.

The answer to the question he poses at the end of the quote above is evidently affirmative. There will clearly be concern on improving students' grades as a way to continue increasing the institution's funding at the end of every year. This effect is, "roughly speaking, [about] the degree to which a test affects the curriculum that is related to it."[23]

As stated above, these three language assessment cornerstones have been granted great importance in the past years and should, no doubt, be made available to every language teacher within and outside the context of Costa Rica's public education system.

Principles of Listening Assessment

This section reviews relevant aspects of listening assessment theory as discussed elsewhere by the authors.[24]

Approaches to Listening Assessment

Buck presents three approaches to the assessment of listening skills. The first is the discrete-point approach which "[breaks] listening into component elements and assesse[s] them separately."[25] The second approach is the integrative approach. According to Oller, "whereas discrete items attempt to test knowledge of language one bit at a time, integrative tests attempt to assess a learner's capacity to use many bits at the same time."[26] Common question types include dictation and cloze. Here, whole language is better than the sum of its parts. The last approach is the communicative approach. Its rationale poses that the listener must be able to comprehend the message and then use it in context. It follows that question formats should be authentic in nature.

23 Brown, 243.
24 Gamboa and Sevilla.
25 G. Buck, *Assessing Listening* (Cambridge: Cambridge University Press, 2001); in Coombe, Folse and Hubley, 91.
26 J. W. Oller, Jr., *Language Tests at School* (London: Longman, 1979).

Types of Listening

Coombe et al.[27] describe two types of listening: general and academic listening. They classify the following micro-skills that are part of general listening: clustering; recognizing redundancy; comprehending reduced forms, hesitations, pauses, false starts, and corrections; understanding colloquial language; processing prosodic features; and understanding and using rules of conversational interaction. Academic listening, on the other hand, includes identifying the purpose and scope of a lecture, the topic, and its logical development; understanding the relationship among discourse units (main versus supporting details); recognizing lexical terms related to the topic; recognizing markers of cohesion (first, next, in conclusion, etc.) and intonation in a lecture, detecting the speaker's attitude toward the subject; and recognizing digressions (turning aside from the main subject) and non-verbal cues of emphasis.

Considerations in Designing Listening Tasks

Teachers must take into account many aspects when they design listening tasks. Before beginning to design a listening test, teachers should consult the course objectives and assessment specifications and guidelines. Tasks should reflect those that occur in real-life situations, and the language used should be natural. In addition, the students should be able to use background knowledge.

The following list of considerations is described by Coombe et al.[28]

Content. Specifications will provide information regarding test content; text types (i.e., narrative, descriptive, etc.); speech types to be used (i.e., phrases, single utterances, two-person dialogues, multi-participant dialogue, monologues); mode of input (audio, video, live reader), varieties of English; scripted or unscripted input; and length of input (in time or number of exchanges).

27 Coombe, Folse and Hubley.
28 Coombe, Folse and Hubley.

Background knowledge. Testers can control background knowledge by writing tasks that exploit specific course materials by providing students with the required background knowledge during testing via advanced organizers or practice prompts. In addition, the primary focus of items should generally be on meaning rather than on form.

Texts. Unavailability of suitable texts is the most pressing issue because creating scripts is not an easy task. Assessment writers should make an inventory of the topics in a course and collect appropriate material in advance. Unfortunately, teachers very often take reading texts and transform them into listening scripts that are unauthentic due to lack of redundant features. Instead, teachers should look for texts and infuse oral characteristics. Use an oral marker at the beginning: "Today I am going to…;" use less complex structures; insert *um, uh, ah;* use *and, but,* or *so* instead of *although, whereas*; read it aloud to make sure it sounds natural, make a script or recording, and include pauses, redundancy, false starts, ungrammaticality, hesitations, etc.

Vocabulary. Students must know 90-95 percent of the words to understand the text/script.[29] When writing a listening test, teachers should include vocabulary from their own word lists into listening scripts whenever possible because lexical overlap can affect difficulty. Teachers must be aware that words used in the passage as well as in the questions and response options when used in the answer key, make the question or answer easier, whereas when used as distractors, the questions or answers become more difficult. Unfamiliar words should never be used as the correct answer.

Test structure. Tests should start with easy questions to reduce test anxiety. They must also test a wide range of skills. Items should be ordered as they are heard. Items should be spaced out. No content

29 P. Nation, *Teaching and Learning Vocabulary* (Boston: Heinle & Heinle, 1990); in Coombe, Folse and Hubley.

from the first 15 to 20 seconds should be tested. Easy as well as challenging items such as paraphrased content and differencing tasks should be included.

Formats. Students should never be exposed to new formats in testing situations. Formats such as multiple-choice questions and true or false items may be used because they are reliable and easy to mark and analyze. Memory plays an important part in listening comprehension tests. More options add to the memory load and affect the difficulty of the task and the question itself.

Item writing. Items should be spaced so that students have time to respond to one item without missing the next. Each new section should be framed with an advanced organizer to help develop the context and activate student's background knowledge.

Timing. Timing will be determined by how many times the students listen to the text. Test-takers should be given the chance to listen to the text twice in achievement tests, but when assessing the main idea the listening passage should be played only once.[30] Finally, students need be given time to pre-read questions.

Skill contamination. Skill contamination refers to the idea that test-takers must use other language skills to answer listening items. Now it is viewed as skill integration.

Research Hypothesis and Methodology

Upon examining current theory on language assessment and listening assessment, it may be asserted that more effective listening assessment practices could be achieved by in-service teachers (from

30 Buck.

the western area of Costa Rica) ranked B1 according to CEF if they are provided with training on listening assessment that draws upon both theory and MEP's testing guidelines.

Research Design

The goal of this study was to examine the incidence of teacher training on listening assessment. It draws upon both theory and MEP guidelines for the listening assessment practices of in-service teachers. The study is quantitative and is based on correlational data analysis techniques, as it examined the relation between two variables, namely teacher training on listening assessment and the tests created after teachers underwent the training by comparing them to tests created by teachers ranked at the same level, from the same area, but who did not take any listening assessment training. Participants in this study were MEP teachers who work in western Costa Rican high schools and were ranked B1 according to the CEF. The participants were divided into two groups, control and experimental. The experimental group was given training on listening assessment via the workshop, while the control group did not receive any training. None of them had received prior in-service training on listening assessment methodology.

Procedure

An eight-hour workshop on listening assessment was designed for the teachers participating in the study, and it was divided into two sessions. In the first session, theory and MEP assessment guidelines were discussed. Participants were presented current testing principles in an interactive fashion; that is, principles were discussed by the researchers, and at the same time participants shared testing experiences as a way to achieve a more solid understanding of the theory. As for MEP assessment guidelines, they were presented in a lecture. In the second session, teachers were asked to create listening tests following the theory discussed in the first session, and the activity was divided into stages. In the first stage, participants put the theory

covered into practice by creating listening tests individually. The tests included a minimum of three tasks, as indicated in the MEP assessment guidelines. During the second stage, participants shared their tests with a peer-reviewer to receive and provide recommendations for further improvement. This was a crucial step since it allowed the test designers to identify weaknesses even in their own tests, because as reviewers they became more critical. In the last stage, a plenary session was held where some participants shared their tests with the rest of the group.

After the tests had been evaluated in the plenary session, they were analyzed using a checklist developed by the trainers to determine the degree of compliance with both MEP assessment guidelines and the theory on listening assessment. Since the authors analyzed tests created by teachers who had not received any in-service training on listening assessment, the results of the analysis in the present study were correlated with those of the earlier ones.

Instrument

The instrument used to assess the teacher-created tests was the checklist mentioned above. In total, it included eleven criteria that sought to assess the degree of compliance of the tests with the theoretical principles discussed in the workshop (see Gamboa and Sevilla for expansion on assessment principles). These criteria included test format, test heading, general test objective, general instructions, credits, balance of item difficulty, specific instructions, listening test techniques, scoring key, face validity, and beneficial backwash.

Data Analysis and Discussion

Tests Created by Teachers in the Control Group

Here the results obtained from evaluating the tests designed by teachers who did not take any in-service training on listening assessment will be discussed. The two criteria where the greatest degree of

achievement was found were test format and test heading. Because compliance with this requirement was below 50%, the tests could be said to be unreliable in both their format and heading.

A low degree of achievement was found in the second group of criteria made up of general test instructions, balance of item difficulty and specific instructions with an average compliance close to 30%. This means that the tests created by teachers who had not received training on listening assessment are not reliable with regard to general test instructions, balance of item difficulty or specific instructions.

A third group of criteria scored even lower on the degree of compliance with test requirements. Credits and listening test techniques complied on an average of 11.6% with these requirements. This low degree of compliance makes these two the next to the lowest criteria in the group highly undermining the reliability of the tests evaluated.

The group of criteria whose degree of compliance with test requirements was the lowest includes general test objectives, scoring key and face validity, their compliance being 0% for the former and 6.25% for the latter. Such low compliance represents null validity regarding these three criteria for the tests analyzed. All of the results can be seen in Table 1.

Table 1: Degree of Compliance with Test Requirements for Tests Created by the Control Group

Criteria	Test Format	Test Heading	Gen. Test Obj.	Gen. Instruct.	Credits	Bal. Item Diff.	Spec. Instruct.	List. Test. Tec.	Scoring Key	Face Validity	Ben. Backwash
T1	100	83.3	0	100	100	100	50	57.1	0	100	100
T2	87.5	91.7	0	100	0	100	50	57.1	0	100	100
T3	87.5	83.3	0	66.7	0	100	75	0	0	0	0
T4	62.5	83.3	0	66.7	0	100	50	0	0	0	0
T5	75	83.3	0	33.3	0	0	100	14.3	0	0	0
T6	62.5	83.3	0	0	100	100	50	0	0	0	0
T7	87.5	66.7	0	66.7	0	100	25	0	0	0	0
T8	87.5	66.7	0	66.7	0	0	50	0	0	0	0
T9	50	66.7	0	66.7	0	0	50	28.6	0	0	0
RS	700	725	0	566.6	200	500	525	171.4	0	100	200
M	43.7	45.3	0	35.4	12.5	31.2	32.8	10.7	0	6.2	12.5

Tests Created by Teachers in the Experimental Group

This section provides details on the results obtained from examining the tests created by teachers after undergoing in-service training on listening assessment via the checklist described above. The two criteria where the greatest degree of achievement was found were face validity and beneficial backwash. Because all of tests fully complied with this requirement, they are highly reliable in both their layout and as a source of feedback for future decision making by curricular authorities.

The second group of criteria where a high degree of achievement was found pertains to test format and test heading, and balance of item difficulty. Roughly speaking, the three criteria depict a degree of compliance of over 80%, which means that after receiving the training, the participants were able to comply satisfactorily with test and test heading required by the MEP and by assessment theory.

General and specific instructions and listening test techniques showed compliance close to 50% with assessment requirements. Being these three sensitive components of assessment instruments, this area must be improved since, as discussed in current assessment theory, it has serious implications for the test taker.

Regarding general test objectives, results depict very low compliance with assessment requirements. This means that, despite having received training, they did not fully internalize the importance of including a general test objective in the evaluation instrument.

For the last two criteria, scoring key and credits, no compliance with assessment principles was observed. In the case of the former, the reasons for the lack of compliance may have to do with the nature of the workshop and the time constrictions that the participants faced when designing the tests. As for the latter, the reasons were that they did not need to give credits because no copyrighted resources were used. Table 2 depicts the individual results for all the tests and the criteria analyzed.

Table 2: Degree of Compliance with Test Requirements for Tests Created by the Experimental Group

Criteria	Test Format	Test Heading	Gen. Test Obj.	Gen. Instruct.	Credits	Bal. Item Diff.	Spec. Instruct.	List. Test. Tec.	Scoring Key	Face Validity	Beneficial Backwash
T1	100	83.3	0	0	0	100	0	57.4	0	100	100
T2	87.5	66.6	0	0	0	100	50	55.1	0	100	100
T3	87.5	91.6	0	83.3	0	100	50	71.4	0	100	100
T4	87.5	83.3	0	0	0	100	50	71.4	0	100	100
T5	75	91.6	0	0	0	0	50	42.8	0	100	100
T6	75	83.3	0	50	0	100	50	57.1	0	100	100
T7	75	83.3	0	66.6	0	100	50	57.1	0	100	100
T8	100	83.3	0	66.6	0	100	25	71.4	0	100	100
T9	100	91.6	0	83.3	0	100	75	71.4	0	100	100
T10	75	83.3	0	86.6	0	100	75	42.8	0	100	100
T11	87.5	91.6	0	100	0	100	100	85.7	0	100	100
T12	75	75	0	83.3	0	0	75	57.1	0	100	100
T13	87.5	100	100	83.3	0	100	75	85.7	0	100	100
T14	87.5	91.6	0	83.3	0	100	50	100	0	100	100
T15	87.5	100	50	83.3	0	100	75	85.7	0	100	100
T16	87.5	83.3	0	83.3	0	100	50	85.7	0	100	100
RS	1375	1383	150	883	0	1400	900	1098	0	1600	1600
M	85.9	86.4	9.3	55.2	0	87.5	56.2	68.6	0	100	100

Conclusions and Recommendations for Further Research

The correlation between teacher training on listening assessment and listening assessment practices after instructors undergo training has been addressed in this study. The conclusions are presented as derived from the analysis of the tests via the checklist to assess the teacher-created tests. The main findings are summarized below.

First, significant improvement can be seen in tests created by teachers who received training on listening assessment methodology whose average compliance with test requirements was 59% per

criterion; this poses a significant contrast with the 20.95% compliance per criterion with test requirements of tests created by teachers who had not received in-service training on listening assessment. The improvements are evident in the criteria of beneficial backwash, face validity, test format, test heading, and listening test techniques. This was a small study that looked at tests created by 9 teachers (20% of the teachers ranked as B1 in the western area) who had not undergone in-service training on listening assessment, and tests created by 16 teachers (representing 37% of the target teacher group) who had taken in-service training on listening assessment. Nevertheless, the results suggest that better listening test design practices could be achieved simply by providing teachers with training on listening assessment. Certainly, this gives insights regarding how desirable teaching practices can be severely undermined by the fact that listening and listening assessment have been a neglected area in language teaching for years.

Second, findings suggest that despite the significant impact made in the test-design practices of MEP teachers, certain aspects need to be reinforced. They include the writing of general and specific instructions, the inclusion of general test objectives, and the improvement of listening test techniques.

Third, the results imply that more evidence is required to measure the impact of teacher-training on the inclusion of the scoring key and the corresponding credits in listening tests. This is due to the time constraints experienced in the workshop described here, and because teachers did not need to provide any credits during the design of their tests in the workshop. Therefore, the researchers suggest that compliance with these requirements should be examined in the future.

Future research should also be oriented towards examining four important areas. First, the listening passage should be studied as a way to obtain a fuller panorama of the assessment practices of MEP teachers in the western area. Second, consideration should be given to the time restrictions experienced in this study, to be able to provide teachers with more listening material to choose from before they design

their tests. This would allow them to select materials that resemble those of "real life." Third, future studies must look at the impact of peer-editing sessions in the improvement of listening tests designed by MEP teachers ranked B1. Finally, similar studies must be carried out with populations ranked in other levels to explore the correlation between language proficiency and listening test design practices.

Appendices

Appendix 1: Workshop Syllabus

Ministerio de Educación Pública
Dirección Regional de Occidente, Departamento de Inglés

General Workshop Information

Workshop Title: Assessment of Listening Skills for Primary and Secondary Education

Duration of the workshop: six hours

Participants' Proficiency Level: B1, according to the Common European Framework of Reference

Instructors: Roy Gamboa Mena and Henry Sevilla Morales

This is an in-service training initiative supported by the *Oficina de la Supervisión de Inglés de la Regional de Occidente del MEP.*

Workshop Description

This is a theoretical-practical workshop for teachers of English from the western area of Costa Rica, with a proficiency level of B1, according to the Common European Framework of Referenced for Languages. The workshop explores both current theory on listening assessment and MEP guidelines for general assessment. It is divided into two sessions. In the first session, listening assessment principles and MEP assessment guidelines will be discussed; and in the second, teachers will design listening tests based on the theory presented during the workshop. The purpose of the workshop is to help in-service MEP teachers to conduct better listening assessment practices within the context of public education.

Workshop Objectives

General objective: The goal of this workshop on listening assessment is to provide MEP teachers ranked B1 with: a) hands-on knowledge of the theoretical principles for listening tests, and b) a focused overview of MEP guidelines for the design of listening assessment instruments.

Specific objectives:

1. to review general assessment theories and principles
2. to examine listening comprehension assessment theories
3. to discuss the principles of listening test task creation
4. to analyze MEP listening assessment guidelines
5. to create tests as a way of putting listening assessment theory into practice

Workshop Methodology

In this workshop, participants are expected to develop effective listening assessment practices upon exploring listening assessment theory and MEP guidelines on assessment. Regarding assessment theory, the participants will be presented current testing principles in an interactive fashion. That is, the principles will be discussed by the presenters but, at the same time, a more solid understanding of them will be achieved through the participants' sharing their own testing experiences. A lecture will be given on MEP assessment guidelines.

Regarding the test design session, it will be carried out in three stages. In the first stage, participants will put the theory discussed into practice by creating listening tests individually. The tests will include a minimum of three tasks, as indicated in MEP assessment guidelines. During the second stage, they will share their tests with a peer-reviewer to give and receive recommendations for further improvement. This is a crucial step since it will allow the test designer

to pinpoint weaknesses in his/her test, as it will allow the reviewers to become more critical of the test they check. In the last stage, a plenary session will be held where some participants will share their tests with the rest of the audience and, together, they will be analyzed to enrich them even further.

Workshop Contents

1- General assessment theories
2- Types of tests
3- Assessment cornerstones/principles
4- Developing assessment
5- Listening comprehension assessment theories
    - General considerations on listening assessment
    - Models of listening
    - Types of listening
6- Considerations on designing listening tasks
    - Content
    - Background knowledge
    - Texts
    - Vocabulary
    - Test structure
    - Formats
    - Item writing
    - Timing
    - Skill contamination
7- Listening test methods
    - Phonemic discrimination
    - Paraphrase recognition
    - Multiple choice questions
    - True or false items
    - Short answer questions
    - Cloze

- Dictation
- Information transfer tasks
- Note-taking

8- MEP and test items

- Construction of objective items
- Construction of production items

Recommended References

Alderson, J. C., and L. Bachman. "Series editors' preface," G. Buck, ed. *Assessing Listening*. Cambridge: CUP, 2001, x.

Brown, D. *Principles of Language Learning and Teaching* (4th ed.). White Plains, NY: Addison Wesley Longman, 2000.

Buck, G. Assessing Listening. Cambridge: CUP, 2001.

Coombe, C., K. Folse, and N. Hubley. *A Practical Guide to Assessing English Language Learners*. Ann Arbor, MI: University of Michigan Press, 2007.

Council of Europe. *Common European Framework of Reference for Languages: Learning, Teaching, and Assessment*. Cambridge: CUP, 2001. Mendelsohn, D. J. *Learning to Listen: A Strategy-Based Approach for the Second Language Learner*. San Diego: Dominie Press, 1994.

Ministerio de Educación Pública. *La prueba escrita.* San José: Departamento de Evaluación, 2011.

Nation, I.S.P. *Teaching and Learning Vocabulary.* Boston, MA: Heinle, 1990.

Nunan, D. "Listening in Language Learning." J.C. Richards and W.A. Renandya (Eds.). *Methodology in Language Teaching: An Anthology of Current Practice*. Cambridge, UK: CUP, 2002.

Oller, J.W., Jr. *Language Tests at School.* London: Longman, 1979.

Osada, N. "Listening Comprehension Research: A Brief Review of the Past Thirty Years," *Dialogue*, 3 (2004), 53-66.

Appendix 2: Listening Test Checklist

Name of participant whose test was assessed: _____

Objective: *To assess MEP teacher-created listening tests for their compliance with theoretical principles of assessment and MEP assessment guidelines and regulations.*

General instructions

1. Assess the listening test by using the checklist below.
2. Read the criterion in the column on the left, and write a check-mark in the column on the right to indicate your assessment.
3. Use explanations and/or examples from the theory and from MEP guidelines and regulations to support your feedback for items marked "Partly" or "No."

Listening Test Checklist

Criteria	Task Achievement		
	Yes	Partly	No
Test Format			
1. Is the layout of the test clear?			
2. Is it suitably and professionally arranged?			
3. Are top, bottom, left and right margins set at 2.5 cm?			
4. Is the typeface style and font size large enough to enable students to read it easily and understand the data included in the test?			
5. Are diagrams, pictures and other test elements well organized?			
6. Is spacing between lines adequate so that the test appears uncluttered?			
7. Are all pages numbered to keep readers oriented on the right sequence of the test?			
8. Are the photocopies clear enough for students to be able to do the exercises?			

Criteria	Task Achievement		
	Yes	Partly	No
Test Heading: Are the following elements included?			
1. the name of the educational institution			
2. the school term and year			
3. the type of test (midterm or final)			
4. data of listening to be tested			
5. the total points and percentage of the test			
6. the school or high school level			
7. a line for the rater's name			
8. a line to write the date when the test will be administered (or the date is already included)			
9. the allotted time for the achievement test			
10. spaces to indicate the points, grade and percentage obtained			
11. a line for the testee to write his/her name			
12. a line for parents to sign the test, if required			
General Test Objectives			
1. Is there an evaluation objective(s) to establish what the testees should be able to demonstrate in regard to their language development?			
2. Is the objective(s) stated clearly, precisely and concisely?			
General Instructions			
1. Is the language focus on what the test takers should do rather than on what they should not do?			
2. Are instructions organized numerically or alphabetically in a proper way?			
3. Are appropriate action verbs used to introduce each set of instructions?			
4. Are explanations and/or examples specific, short and clear?			
5. Is important information highlighted when necessary?			
6. Is language adjusted appropriately for the students' English level?			
Credits			
Are copyright laws followed by giving credit to the authors of intellectual works such as stories, poems, illustrations, maps, and others?			

Criteria	Task Achievement		
	Yes	Partly	No
Balance of Item Difficulty			
Is the test arranged from the easiest to the most difficult tasks?			
Specific Instructions			
1. Are explanations specific, short and clear?			
2. Is there sufficient context for the test task to be carried out correctly?			
3. Is the language adjusted appropriately to meet the students' level of English?			
4. Is the total number of points and individual value of each correct item included?			
Listening Test Techniques			
1. Are there appropriate test techniques to elicit those behaviors that reflect the students' specific listening abilities more reliably?			
2. Is there a minimum of three different exercises to evaluate specific listening skills?			
3. Are the questions ordered in the same way as the content is heard in the passage?			
4. Are the questions spaced out in the passage?			
5. Is each new section is framed with an advanced organizer to help activate the testee's schemata?			
6. Do the test tasks reflect real-life situations?			
7. Are the items spaced far enough apart so testees have enough time to answer one item without missing the next?			
Scoring Key			
Is there a scoring key specifying the acceptable answers for all listening test items?			
Face Validity			
Do the test content and tasks meet the objectives intended by the test designer?			
Beneficial Backwash			
Do the test content and techniques correspond to the objectives of the curriculum for which this achievement test is intended, so that its eventual administration may have a positive impact on the testees?			