FINDING PROPER PRONUNCIATION ASSESSMENT

HO DANG TUONG NGUYEN

Ho Chi Minh City Open University, Vietnam
Email: nguyen.hdt@ou.edu.vn

(Received: November 14, 2018; Revised: December 05, 2018; Accepted: December 6, 2018)

ABSTRACT

Pronunciation is an indispensable component of a student's language competence. However, there has been a lack of a system to help teachers conduct proper assessment or design reliable tests to evaluate their student's articulation performance in English. This paper presents the atomistic and holistic testing with the aim of contributing to the design of reliable and valid pronunciation tests for English learners.

Keywords: Assessment; Pronunciation; Reliability; Testing; Validity.

1. Introduction

In the era of the World Industrial Revolution 4.0, English has constantly been being used widely around the world, which led to the issue of setting an appropriate pronunciation standard in second language listening and speaking tests (Canagarajah, 2006; Elder & Davies, 2006; Jenkins, 2006; Isaacs & Trofimovich, 2017). Nevertheless, there is limited work that tends to target language testers in this paradigm. Besides, for test efficiency, validity and reliability need to be present. As these two conditions are important for the effectiveness of testing, it is generally accepted that a precise evaluation of students can be achieved if they are both consistent. This paper, hence, aims at finding feasible framework for designing a proper pronunciation test and means of quality control in the test production.

Teaching a language or any subject is obviously connected with formal or informal feedback, making assessment, and giving recommendations for improvement. Giving feedback on pronunciation is not an exception either, which is important for maintaining the student's motivation, providing them with information on how they are progressing and what they should focus on. Compared to learning grammar or vocabulary, where students are able to assess themselves objectively having the correct answers at their disposal, self-evaluation in pronunciation is complicated since it is very often distorted by the student's own ear and phonetic ability. For this reason, the importance of teacher's role in giving feedback on this language component is crucial. Yet, assessing pronunciation, in comparison with other language components seems to be a difficult task when not much attention is paid to this issue in the literature. The main reason for this negligence is the fact that "speaking, comprising pronunciation, is a skill which is too complex to enable any reliable analysis which could be considered to be objective" (Sebestova, 2007, p. 17).

2. Assessment methods

Learning pronunciation is based on both production and reception; therefore, it is closely connected with oral exams and listening tests. Contrary to production, reception can be tested and quantified by counting the correct answers in a written test, whereas production is more difficult to assess as it involves testing speaking or reading which, from the listener's point of view, include many other aspects of communication, not pure sounds (Sebestova, 2007). It is always subjective to a large extent, and so the problematic matter of assessing pronunciation production is its reliability. We are bound to rely on the impression of the assessor. Therefore, both the procedure and the assessment should be defined as concretely as possible. Depending on the aim of the assessment, two approaches of testing pronunciation are described – atomistic and holistic.

2.1. Atomistic testing

Atomistic approach requires "a detailed marking scheme in which specific aspects of pronunciation are evaluated separately" (Sebestova, 2007, p. 19). It means reading aloud word lists based on phonemic oppositions, short sentences containing minimal pairs or it also enables testing appropriate sentence accentuation or sentence stress and intonation. This approach is claimed to be more objective than the holistic one as it judges only segments of speech – particular vowels, consonants, stress, rhythm, intonation, etc. Nevertheless, the drawbacks of this approach are the demands on the assessor. It requires recording the learners' speech samples and repeated listening to them, so it is extremely time-consuming and thus unsuitable for large classes.

2.1.1. Repetition

At the beginner level, the easiest test to prepare is repetition exercise. It is useful for learners who cannot read or who are beginning with English. It is based on hearing sounds, stress and intonation, and imitation which gives the teacher the gist of learners' potential and phonetic ability. The test may consist of single words or sentences checking particular items rather than all pronunciation aspects at the same time in order to be as much objective as possible (Sebestova, 2007).

2.1.2. Hearing identification

Another way of testing "beginners as well as more advanced learners" is hearing identification (Madsen, 1983, p. 61). Good pronunciation is dependent on our ability to hear the language. It can consist of recognizing sounds in minimal pairs, the fall or rise in intonation or identifying stress in words or sentences.

2.1.3. Reading aloud

Commonly used way of pronunciation assessment is reading aloud. According to Madsen (1983) three points should be kept in mind: (1) When using lists of sentences, evaluate only one or two points per sentence; (2) use natural language; [and] (3) avoid signalling to the student which pronunciation point you are testing (p. 66).

Since reading tends to be longer and involves many points to assess at the same time, it is advisable to record the learners' performances in order to listen to them repeatedly and have the possibility to compare. The material to read should enable natural sound, e. g. a letter, instructions etc., and students should have time to read the text silently before reading for assessment to get the context. The reading aloud testing provides good control and enables to test almost all aspects of pronunciation including stress and intonation as well as vowels and consonants. Nevertheless, we have to count on the fact that reading and speaking skills are not the same and, inevitably, the intonation and sentence rhythm in reading is usually not as natural as in normal conversation.

2.2. Holistic testing

A higher level of achievement is testing the "intelligibility and acceptability of the learner's performance" (Sebestova, 2007, p.21). In this holistic approach to pronunciation testing, "examiners are asked not to pay too much attention to any one aspect of a candidate's performance, but rather to judge its overall effectiveness." (Alderson, Wall & Claphaim, 1996, p. 289).

The advantage of this procedure is that it can be administered to large groups and is not as time-demanding as the atomistic approach. This approach is used in many international exams in English, where the pronunciation is involved in so-called intelligibility and acceptability of the candidate's speaking performance. The definition of intelligibility is rather general: "Intelligibility is being understood by a listener at a given time in a given situation. So it's the same as "understandability", which means "The more words a listener is able to identify accurately when said by a particular speaker, the more intelligible that speaker is" (Kenworthy, 1987, p. 13). The issue of intelligibility is complex and is a major part of communication. Therefore, the goal is not only the correct production of sounds, stress patterns and intonation, but efficiency of communication without irritation and difficulties understanding. So the goal of pronunciation can be defined as "comfortable intelligibility" (Sebestova, 2007, p. 23).

The main criterion for holistic testing is the efficiency of communication between two people. Therefore, the best method is interactive testing including more than only one participant (Sebestova, 2007). All the activities should be used in the interaction of the assessor or another student to involve both sides of the communication – the speaker and the listener – to function as an oral interview including natural situations and asking questions. Below are some recommendations by Sebestova (2007, pp. 24-25).

2.2.1. Re-telling stories

This kind of test involves first reading a story silently and then telling the story using one's own words and sentence structures. The assessor may interfere giving further questions.

2.2.2. Description of pictures

Pictures may be used for description of objects, people or scenes, or for comparison of two similar pictures, in which the learner looks for similarities and differences.

2.2.3. Sequence of pictures

This test is based on telling a story involving linking words expressing the cause and the result. It can be applied to only one student or a pair where each of them is given one-half of the pictures and they should decide on the correct sequence of the story.

2.2.4. Pictures with speech bubbles

In this test, students are required to guess what the people in the pictures are saying. It may be used individually or in pairs.

2.2.5. Using maps

Many student's books involve a unit dealing with giving directions. This activity is to be done in pairs, where one gives the directions and the other one follows them.

3. Means of quality control in test production

In the design of any assessment instrument test, developers must be concerned with identifying potential sources of error in the instrument and providing evidence to justify test score interpretations. These two aspects are addressed and discussed as reliability and validity respectively.

3.1. Validity

According to Owen (1997, p. 13), two areas should be considered when discussing validity in testing: (1) consider how closely the test performance resembles the performance we expect outside the test, and (2) consider to what extent evidence of knowledge about the language can be taken as evidence of proficiency (p. 13). Referring to the importance of validity in tests, Cohen et al. (2000) state that effective research is impossible or even "worthless" without the presence of validity (p. 105), though they do recommend against aiming for absolute validity. Instead, they define the search for validity as being one of minimizing invalidity, maximizing validity, and therefore using measurement in validity as a matter of degree rather than a pursuit of perfection (p. 105).

Several categories exist for validity. The following four categories are described by Hughes (1989) and Bachman (1990), these being construct validity, content validity (included within this are internal and external validity), criterion-based validity, and face validity.

3.1.1. Construct validity

Construct validity is concerned with the level of accuracy a construct within a test is believed to measure (Brown, 1994, p. 256; Bachman & Palmer, 1996), and particularly in ethnographic research, “must demonstrate that the categories that the researchers are using are meaningful to the participants themselves” (Cohen et al., 2000, p. 110).

3.1.2. Content validity

Content validity is concerned with the degree to which the components of a test relate to the real-life situation they are attempting to replicate (Hughes, 1989, p. 22; Bachman, 1990, p. 306) and is relevant to the degree to which it proportionately represents. Within the domain of content, validity includes internal validity and external validity. These refer to relationships between independent and dependent variables when experiments are conducted. External validity occurs when our findings can be related to the general populous, whereas internal validity is related to the elimination of difficult variables within studies.

3.1.3. Criterion-related validity

Criterion-related validity “[relates] the results of one particular instrument to another external criterion” (Cohen et al., 2000, p. 111). It contains two primary forms, these being predictive and concurrent validity. Concerning predictive validity, if results from two separate but related experiments or tests produce similar results, the original examination is said to have strong predictive validity. Concurrent validity is similar, but it is not necessary to have been measured over a span of time and can be “demonstrated simultaneously with another instrument” (Cohen et al., 2000, p. 112).

3.1.4. Face validity

This term relates to what degree a test is perceived to be doing what it is supposed to. In general, face validity in testing describes the look of the test as opposed to whether the test is proved to work or not (Nadasdy, n.d.).

3.2. Reliability

Reliability relates to the generalisability, consistency, and stability of a test. Following on from test validity, Hughes (1989) points out that “if a test is not reliable, it cannot be valid” (p. 34). Hughes (1989) continues that “to be valid a test must provide consistently accurate measurements” (p. 50). Therefore it would seem that the higher amount of similarity there is between tests, the more reliable they would appear to be (Hughes, 1989). However, Bachman (1990) argues that although the similarity case is relevant, other factors concerning what we are measuring will affect test reliability. Factors including test participants’ personal characteristics i.e. age, gender, and factors regarding the test environment and condition of the participants can contribute to whether or not a test is effectively reliable (p. 164).

Terms relating to methods estimating reliability include inter-rater reliability and test-retest reliability. These methods each have their own ways of examining the source of error in testing. Inter-rater reliability is concerned with how scores from various sources are balanced and importantly to what degree markers scores are showing equality (Nunan, 1992, pp. 14-15). Test-retest reliability gives an indication as to how a test consistently measures individual performances of students that are tested across various testing organizations (Underhill, 1987, p. 9). A further simplified definition is offered by Nunan and Weir and Roberts stating that inter-rater reliability is the degree to which the scores from two or more markers agree (Nunan, 1992, pp. 14-15; Weir & Roberts, 1994, p. 172).

3.3. Ensuring validity

Hughes (1989) states that the concept of test validity can seem uncomplicated but on closer inspection can appear highly complex (p. 34). Some experts say that "one might suppose that ultimately there is no means of knowing whether a test is valid or not" (Owen, 1997, p. 13). One certainty is that it is possible to describe and assess test validity in various ways. Initially, one could attest that the most important description is based on test effectiveness. Hughes (1989) points out the basis for a simple criterion for test quality and offers evidence for showing relevance of certain descriptions that may help to rectify difficulties in language testing.

Firstly, Hughes (1989) states specifically that a test should simply "... [measure] accurately what it is intended to measure" (p. 26) to assure us of its validity. Though this may appear relatively simple in terms of straightforward testing, several definitions of what we expect our students to achieve can overcomplicate what we are attempting to measure. To assist in simplifying ambiguous "theoretical constructs" such as fluency in speaking, reading ability, etc. certain descriptions of validity can be considered including construct validity, content validity, and criterion-related validity. The following considers these variants. With content validity, Hughes (1989) points out that if the test has positive content validity it is more likely to accurately test what is required, and thus leads to constructing validity. He states that "the greater a tests content validity, the more likely it is to bean accurate measure of what it is supposed to measure" (1989, p. 27). Importantly, when creating tests, specifications have to be established at an early stage referring to what is required from the tests participants. These specifications should be areas that are considered to be of maximum benefit when defining that which is to be measured and achieved through the testing. Hughes (1989) purports though that "too often the content of tests is determined by what is easy to test rather than what is important to test" (p. 27). Therefore it is important to be clear about what is required. Criterion-related validity provides assessment from different perspectives and presents an opportunity to compare qualitative score analysis against quantitative independent judgments oftest participants'abilities. Hughes (1989) states that all of these "have a part to play in the development of a test" (p. 30).

Hughes (1989) also draws our attention to how scoring is important when judging the validity of tests and how testers and test designers must "make sure that the scoring of responses relates directly to what is being tested" (p. 34). Accurate scoring of responses would seem imperative if correct measurement is to be assured. Being clear as to what is required as a response, e.g. clear responses of pronunciation on speaking tests should not be confused with hesitation or intonation issues, validity may then be more achievable and measurements more accurate and relevant.

3.4. Ensuring reliability

According to Hughes (1989, pp. 44-50), there are several ways to ensure reliability. These include gathering information about the test candidate by adding extra and more detailed questions, tasks, and examples to tests, balancing the difficulty of questions so they do not "discriminate between weaker and stronger students", focusing and restricting questions that may allow for too much elaboration, avoiding ambiguous questions and items, being clear with instructions for tasks, presenting tests clearly to avoid confusion, practicing the test format with students so that they are familiar and prepared for the actual test, encouraging consistency across administrations on large scale testing, using items that utilize objective scoring i.e. providing part of an answer for a test taker to complete rather than eliciting an entire sentence as an answer, restricting the freedom afforded to candidates in terms of the comparisons made between them, providing clear and detailed score keys, helping testers and scorers by training them at an early stage and conferring with test designers and testers about how responses are to be scored before scoring commences, having students represented by numbers rather than personal details to restrict any possible bias occurring, and using, if possible, independent scorers to evaluate objectively eliminate discrepancies.

Though the variable in human errors in testing between testers and candidates are significant, these items seem to at the very least work towards creating better reliability (Nadasdy, n.d.). It would certainly seem of benefit to have practical experience of teaching and testing enabling researchers a firsthand experience of what may be required throughout the entire process of test organization.

In short, to design a proper pronunciation test and assure its validity and reliability, a teacher or an instructor or a test designer could refer to the framework below.

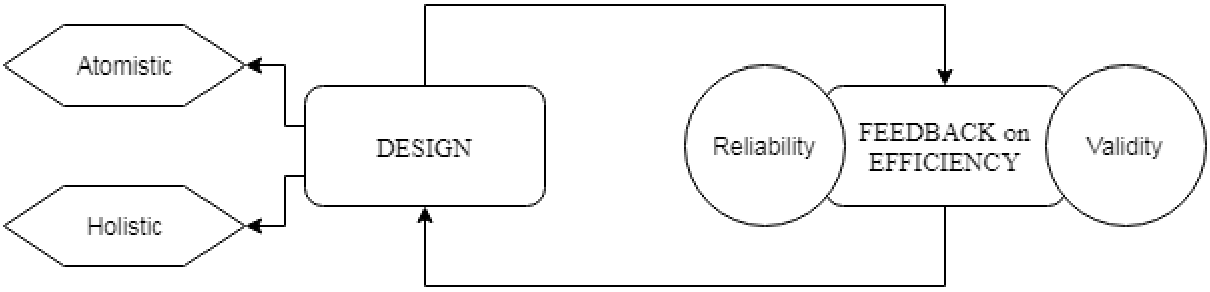

Figure 1. A suggested framework for designing pronunciation tests

4. Conclusion

To conclude, both holistic and atomistic testing (except the method of hearing identification) are suitable for the oral assessment of students' pronunciation. Which one is more suitable depends on the purpose of testing. As far as reliability is concerned, atomistic tests are more reliable for diagnostic purposes in the language classroom and in cases in which scoring is carried out by different assessors, whereas holistic approach is faster and more appropriate for experienced assessors (Hughes, 1989)■

References

Alderson, C. J., Wall, D. & Claphaim, C. (1996). *Language Test Construction and Evaluation*. Cambridge, UK: Cambridge University Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. UK: Oxford University Press.

Bachman, L. & Palmer, A. (1996). *Language Testing in Practice*. Oxford University Press.

Brown, H. D. (1994). *Principles of Language Learning and Teaching* (3rd ed). Prentice Hall.

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an International Language. *Language Assessment Quarterly*, *3*(3), 229–242.

Cohen, L., Manion, L. & Morrison, K. (2000). *Research Methods in Education*. Routledge.

Elder, C. & Davies, A. (2006). Assessing English as a lingua franca. *Annual Review of Applied Linguistics, 26(*1), 232–301.

Hughes, A. (1989). *Testing for Language Teachers* (1st ed). Cambridge, UK: Cambridge University Press.

Isaacs, T., & Trofimovich, P. (2017). *Second Language Pronunciation Assessment: Interdisciplinary Perspectives*. Bristol, UK: Multilingual Matters.

Jenkins, J. (2006). The spread of EIL: A testing time for testers. *ELT Journal, 60*(1), 42–50.

Kenworthy, J. (1987). *Teaching English Pronunciation*. Harlow: Longman.

Madsen, S. H. (1983). *Techniques in Testing*. New York, USA: Oxford University Press.

Nadasdy, P. B. (n.d.). Reliability and validity of a test and its procedure conducted at a Japanese high school. Retrieved from http://www.nuis.ac.jp/ic/library/kiyou/14_nadasdy.pdf

Nunan, D. (1992). *Research methods in language learning*. Cambridge University Press.

Owen, C. (1997). *Testing*. Birmingham, UK: The Centre for English Language Studies.

Sebestova, S. (2007). *Some aspects of assessing pronunciation in EFL classes* (Diploma thesis, Masaryk University, Brno). Retrieved from Google Scholar.

Underhill, N. (1987). *Testing Spoken Language: A handbook of oral testing techniques*. Cambridge University Press.

Weir, C. & Roberts, J. (1994). *Evaluation in ELT*. Blackwell Publishing.