Pedagogical Quality of English Achievement Tests: An Untold Story of Iranian High School Students' Oral Scores

Hossein Heidari Tabrizi [1*]

[1]Associate Professor, Department of English, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran

Available online

Keywords:

English Achievement Tests, Iranian High Schools, Language Assessing Quality, Students' Oral Scores

Abstract

The test scores on students' report cards is the only benchmark against which their English achievement are assessed for evaluative purposes in Iranian high schools. According to the rules and regulations, the average score of a student is required to be reported by English teachers, of course based on her/his performance on both oral and written tests. The present study aimed to discover if Iranian high school female students' English scores on their report cards represent the real sum of their oral and written test scores. To do so, the average scores of 30 female students in Grade 11 at two Iranian girls' senior high schools in Isfahan were compared with those of a researcher-made validated oral and written test. The results showed that the scores of the students on the newly developed test were higher than those recorded on their report cards. The results of a paired *t* test revealed a statistically significant difference between the means of these two sets of scores rejecting the common false presupposition about students' low performance in oral skills. Teachers typically skipped the oral test and rated their students' oral ability just based on their own intuition or students' performance on the written test. It seems that the exclusion of the oral test leads to this difference in the scores. In conclusion, Iranian high school students' English scores appearing on their report card are not a sound reflection of their performance on the oral and written tests.

Introduction

Assessment is one of the fundamental bases for teaching English as second or foreign language. Though not always used and interpreted as positive, assessment has always been an indispensable part of school life in general and English classes in particular. The new millennium has however witnessed a drastic change in classroom assessment: the quantity, the quality, the variety and the diversity of assessment practices have increased dramatically (Bachman, & Palmer, 2010; Brown, 2013; Brown & Abeywickrama, 2010; Douglas, 2010, Purpura & Turner, 2017; to name but a few).

More challenging is assessing the so-called 'speaking skills'; that is, spoken-language proficiency in language teaching. "When a foreign language is an assessed school subject, the procedures used to arrive at a valid account of learners' ability to speak is of crucial importance to schools, teachers and learners" (East, 2016, p. v). In an English as a Foreign Language (EFL) educational environment, exposure to spoken language is either minimal (if not zero) outside formal instruction settings or just limited to the instruction period in English classes. "However, the prevalence of local languages as the medium of communication has reduced the scope for cultivating good spoken English among the subjects being studied" (Jayaraman, 2017, p. 134).

A proficiency test as part of a high-stakes school-based assessment is very different from language used outside of the classroom in a naturalist context for communication (Bachman, 1990). It is challenging for proficiency tests not only to replicate but also to capture authentic language use. Changes in assessment practices – a move from assessment *of* learning towards assessment *for* learning (Fulcher & Davidson, 2012; Tsagari & Banerjee, 2016) – may not align with learners' expectation about the nature and function of assessment. If teachers move away from summative tests and adopt continuous assessment based on an assessment *for* learning approach (Brown & Hudson, 2002), there may be resistance from learners as what counts for them may be their end of course scores rather than the teacher's well-intentioned philosophy of assessment.

As Brown (2002, p.14) states, washback become negative when there is a mismatch between the content that is the material and abilities being taught, and the test, in fact much time of the class is always spent just on the materials that come to test. In fact, standardized achievement tests are considered helpful because they rank students based on what students know and can do but not

because they successfully measure how well those students have been taught. However, in the era of accountability, schools are perceived as better or worse based on their proficiency, readiness, or growth; and teachers are believed to be more effective when their students perform better on high-stakes achievement assessments (Court, 2010; Popham, 1995).

As a result of the centralized system of education in Iran, English high school teachers are moved to raise their students' achievement test scores so that the scores will look good on their school report card. As Ghorbani, Arshad, Sahandri, and Nooreen (2008) correctly argue, test scores provide the only criterion against which the students' progress at school are indicated in Iranian high schools. Likewise, the only benchmark against which the achievement of schools is evaluated is the test scores and the GPAs of the students. Thus, "test curriculum alignment and teaching to the test have influenced the meaning of scores. The focus of educational accountability has increased pressure to raise test scores in Iranian high schools" (p. 145). According to the rules and regulations enforced by the Ministry of Education, teachers have to sum the oral and written test scores, and record their resulting average scores on the final report card of the students.

However, it seemed to the researcher, at least based on common sense and experience, as well as based on the previous research studied conducted (Abbasi, Heidari Tabrizi, & Chalak, 2021, 2020; Abedini & Chalak, 2017; Aslani & Heidari Tabrizi, 2015; Heidari Tabrizi, & Onvani, 2018; Mizbani & Chalak, 2017; Zafari, Heidari Tabrizi, & Chalak, 2020) that many achievement tests in Iran failed to reflect effectively whether teachers successfully covered and delivered the necessary content in their instruction and assessment. One piece of evidence can be the frequent negative feedbacks teachers are likely to receive from the students about the final tests every academic year. Still another piece of supporting evidence is the countless anecdotes one hears in professional conferences about the deficiencies of these tests. As a matter of fact, it seems that teachers prefer to ignore the oral test to avoid the possible score decrease. Accordingly, the present research aimed to find answer to the following question:

RQ. Do Iranian high school female students' English scores on their report cards represent the real sum of their oral and written test scores?

Methodology

Research Design & Setting

To achieve the research objectives, a comparative descriptive quantitative research design was employed. The study was conducted in a state top-ranked girls' high school in Isfahan, Iran during the academic year 2019-2020.

Participants

Due to the access limitation, the participants in the present study were selected non-randomly through non-probabilistic availability sampling. In practice, all Grade-11 students (N=30) studying at two homogeneous equal classes at *Narjes Tohid Girls' High School* in Isfahan in the academic year 2019-2020 attended the study as the subjects. The educational system under the jurisdiction of Iran Ministry of Education currently in use can be illustrated in terms of level, duration and age ranges in Table 1:

Table 1

Iran's Educational System

	Level		Grade	Duration	Age Range
1	Elementary School		1-6	6	6-11
2	High School	Junior	7-9	3	12-14
		Senior	10-12	3	14-17

As evident from Table 1, the subjects in this study were all 16-to-17-year-old females in their penultimate year of senior high school.

Instructional Materials

The English textbook currently used in Iranian high schools; namely, *English Book 2* served as the instructional materials in the classes under investigation and the content of the final written achievement test as well as the researcher-made oral proficiency test

were taken from this very textbook. The textbook series is locally designed to cater for and respond to the English language needs of Iranian students in Grade 11.

Measuring Instruments

In the present study, two instruments were employed to evaluate the students' attainment: an official final written standard achievement test and a set of researcher-made test items in form of an oral interview.

Official Final Written Test

The first instrument of the study was actually the very official final written standard achievement test already developed by the test designers in the Isfahan Department of Education for *English for Grade 11*. In fact, the test was the routine standard one administered by the school at the end of the instruction and claimed to enjoy good reliability and validity. In other words, the test was designed based on the textbook contents and the regulations of the Department and the rules enforced by Iran's ministry of Education. Table 2 summarizes the test specifications in terms of language skills and components as well as their weight, and test form used:

Table 2

Item Type Distribution and Item Weight in Final Written Test

Item	Question Type(s)	Score
Vocabulary	Fill in the blanks	2
	Multiple choice questions	2
Comprehension	Completion, Multiple choice items, True/False, Wh. questions, Yes/No questions	4
	Sentence Comprehension	3
Dictation	Spelling completion	3
Grammar	Multiple choice items	2
	Scrambled sentences	2
	Sentence completion	2
	Answering questions via pictures	2

Language Functions	Completing conversations	3
Pronunciation	Finding the different words, Putting the words under the correct column	1
Substitution Items	Substituting words in sentences	4
Total Score	Test items overall	30

Oral Proficiency Test

To evaluate the oral proficiency of the students in English, the researcher developed a set of test items to be used in an interview session. To construct a valid, reliable user-friendly instrument, the researcher followed the principles proposed by East (2016) as well as the guidelines recommended by Luoma (2004). To establish the construct validity as well as the content validity of the instrument, the researcher followed the rules and guidelines officially notified by

Iran's Ministry of Education.

Having been designed and prepared, the test items were studied by five experienced teachers and test developers who confirmed their validity based on experts' opinion. A pilot study showed that the instrument enjoyed a relatively high reliability (α=.75). To enhance the reliability of the instrument in general and its inter-rater reliability in particular, the performance of the students during the oral proficiency test session was rated and scored by three experienced English teachers.

The scoring was done in practice based on the rubric enforced by Iran's Ministry of Education for evaluating students' performance as far as their oral proficiency in English is concerned. According to O'Sullivan, "The scoring system includes everything that is done to transform a test performance into a test score" (2014, p. 168). Table 3 presents the scoring rubric and its different domains and their relative weight:

Table 3

Test Rubric for Evaluating Students' Performance on Oral Proficiency Test

Item	Question Type(s)	Score
Reading Aloud	Speed	
		1
	Comprehension	**1**
	Fluency	**2**
	Rhythm, Stress, Pronunciation	**2**
Text Reproduction	Retelling stories	**2**
Language Functions	Conversations & Dialogs	**1**
Pronunciation	Correct pronunciations	**1**
Total Score	Test items overall	**10**

Research Procedures

To collect the data, first in February 2017 the official written final test of the course *English for Grade 11* was administered in the target high school and the papers were scored by the English instructor of the two classes. Afterwards, the researcher had a short talk with the instructor to find out what her opinion was about students' oral test. It was revealed that like the other English high school teachers, she rated the students' oral ability intuitively based on their performance on the written final test.

Later, the researcher asked her to provide him with the students' written scores as well as the scores recorded on their report cards as the total final scores for the course. Subsequently, to assess their oral ability in English, the researcher asked the students to participate in an oral 3-to-5-minute interview a week after their final test while their voices were recorded. To eliminate the students' possible fear and to increase the reliability of the obtained scores, they were informed that the results of the study would be kept confidential through anonymity procedure and would have no effects whatsoever on their final scores recorded in their report cards beforehand. Three teachers with more than 10 years of experience in teaching and assessing English oral skills were asked to score the students' performances from 0 to 10 using the rubric for the oral test suggested by the Ministry of Education.

Finally, the students' oral scores calculated by the researcher were added to the scores they received on the written final test. The next step taken was to compare these scores with those already recorded in their report card. To conduct such a comparative statistical analysis, the researcher, due to the nature of the interval data collected, made use of descriptive as well as inferential statistics.

Results

Having collected the data in terms of the students' obtained English scores on the oral test as well as the report card, the researcher first summarized the whole interval data using measures of central tendency and variability. Table 4 presents the descriptive statistics for these scores.

Table 4

Descriptive Statistics for Students' Score on the Report Cards and the Oral Test

Test	N	Min	Max	Mean	SD
Standard Test	30	15	20	18.31	1.77
Researcher-made Oral-Written Test	30	16	20	18.9	1.39

As shown in Table 4, the mean score for the researcher-made oral-written test was higher than that of the scores recorded in student report cards based on the written standard test. In fact, completely contrary to the expectations and presupposition, students showed relatively high proficiency in oral skills. It was found out that most of them showed higher ability in this test than it was supposed to be. Figure 1 illustrates the scores:

Figure 1

Graphic representation of the means for final scores of students' written & oral score
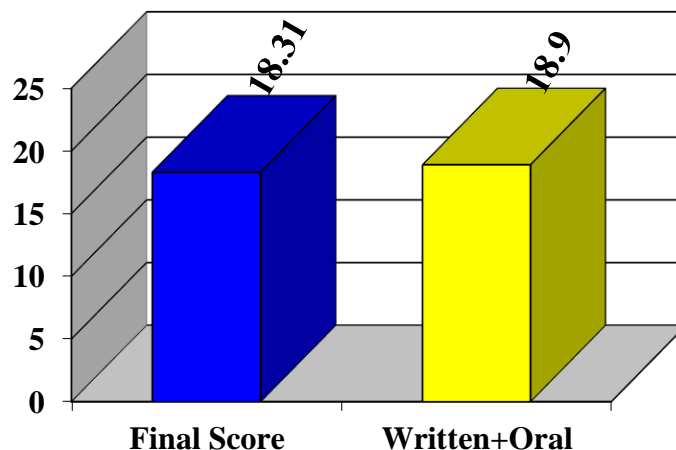
Next, to see whether there was a statistically significant difference between the two means of these two sets of scores, a matched *t* test was run. The results revealed that the observed *t* value (5.813) was much greater than the critical *t* (3.396) at an optimal probability level (p<0.01). Thus, the possibility of any chance differences between the two can safely rejected. In other words, the score of students increased in a meaningful way as a result of actually administering an oral test to them.

Discussion and Conclusion

Evidently, it can be inferred that written exams cannot measure students' oral proficiency at all and an oral test is required to score students precisely. Therefore, it can be concluded that an oral test must be taken at high schools besides the written tests to determine students' oral proficiency in English to have a more accurately made evaluation and decision. One possible interpretation can be stated as follows. English teachers do not usually pay much attention to the oral test and score students just based on their performance in written test because students are not tested orally at the provincial or national level. The results of the present study are in broad agreement with those of Jahangard (2007), who maintained that students' aural and oral skills are not emphasized in Iranian high school in prescribed EFL textbooks.

As the results showed, Iranian high school students' overall scores are not actually the real sum of their oral and written exam. English teachers neglect the oral exam and rate students only based

on their performance on the written exam. Because of this wrong presupposition about their oral ability, students are not tested orally in provincial and national level; thus, the teachers as well as the students do not pay attention to English oral proficiency especially speaking skills. The English teacher participated in this research study believed that the students do not have enough oral ability in English, and if the teacher tries to take any oral test, a considerable decrease will actually be observed in the scores of the students, and as the oral parts of the text books are not tested at the provincial and national level; teachers narrow down the content of text books only in written test. They neglect any oral exam and rate their students just based on their written performance.

Moreover, teachers prefer to skip what they think might have a negative effect on the scores. Then the real fault for inappropriate forms of scoring procedure lies not so much with teachers, but with policymakers who have created accountability systems centered on higher test scores, with little regard to students' real ability and the way scores are attained. Instead of use test result to panelize schools, tests should become stronger tools for improving students, schools' principals, and teachers.

By way of conclusion, it is worth of mentioning that teachers do not pay enough attention to this fact that their students especially in the recent decade have changed a lot. They have easy access to internet, and they spend so much time in application and websites such as *Facebook, Telegram* or *Instagram*. In addition, many students spend more time in this social media more than any other entertainment or environments. On the other hand, parents in recent years push their children to learn English as a foreign language. In some cases, parents even force their children to learn it. Thus, it seems logical that students' oral ability has changed in comparison of for example ten years ago, while teachers' presupposition about students' oral ability has not changed. In fact, their wrong presupposition about students' lack of oral ability lead to neglecting oral exam and weighty deviation from students' true scores. In practice, teachers prefer to skip the oral test in order for the students to have higher scores while taking a true oral exam might result in an increase in students' scores.

References

Abbasi, S., Chalak, A., & Heidari Tabrizi, H. (2021). Impact of online strategies-based instruction on Iranian advanced EFL learners' speaking scores. *International Journal of Foreign Language Teaching & Research, 9*(36), 21-37.

Abbasi, S., Chalak, A., & Heidari Tabrizi, H. (2020). Effect of online strategies-based instruction on Iranian EFL learners' speaking scores: a case of affective and social strategies instruction. *Journal of Modern Research in English Language Studies, 9*(36), 21-37.

Abedini, F. & Chalak, A. (2017). Investigating the inhibitive factors in the speaking of Iranian EFL learners. *Journal of Applied Linguistics and Language Research, 4*(6), 82-97.

Al-Mahrooqi, R., Coombe, Ch., Al-Maamari, F., & Thakur, V. (Eds.). (2017). *Revisiting EFL assessment: Critical perspectives*. Singapore: Springer.

Aslani, S. M., & Heidari Tabrizi, H. (2015). Teaching Grammar to Iranian EFL Learners through Blended Learning Using Multimedia Softwares. *Journal of Applied Linguistics and Language Research, 2*(8), 76-87.

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

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Brown, J. D. (Ed.). (2013). *New ways of classroom assessment* (Rev. ed.). Alexandria, VA: TESOL.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). White Plains, NY: Pearson Education.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge, England: Cambridge University Press.

Cheng, L., & Fox, J. (Eds.). (2017). *Assessment in the language classroom*. London: Palgrave.

Court, S. C. (2010). *Instructional Sensitivity of Accountability Tests: Recent Refinements in Detecting Insensitive Items*. Paper presented at the Council of Chief State School Officers' National Conference on Student Assessment, Detroit, MI.

Douglas, D. (2010). *Understanding language testing*. New York: Routledge.

East, M. (2016). *Assessing foreign language students' spoken proficiency: Stakeholder perspectives on assessment innovation*. Singapore: Springer.

Fulcher, G., & Davidson, F. (2012). *The Routledge handbook of language testing*. New York: Routledge.

Ghorbani, M. R., Arshad A. S., Sahandri, M., & Nooreen, N. (2008). All that glitters is not gold: Curriculum alignment and improving students' test scores. *International Iranian Journal of Language Studies, 2*, (1), 19-40.

Heidari Tabrizi, H., & Onvani, N. (2018). The impact of employing telegram app on Iranian EFL beginners' vocabulary teaching and learning. *Applied Research on English Language, 6*(4), 1-18.

Hidri, S. (Ed.). (2020). *Changing language assessment: New dimensions, new challenges*. Cham: Springer.

Hidri, S. (Ed.). (2021). *Perspectives on language assessment literacy: Challenges for improved student learning*. London: Taylor & Francis.

Jahangard, A. (2007). Evaluation of EFL materials taught at Iranian public high schools. *The Asian EFL Journal, 9*(2), 130-150.

Jayaraman, S. (2017). EFL assessment: Assessment of speaking and listening. In R. Al-Mahrooqi, Ch. Coombe, F. Al-Maamari & V. Thakur (Eds.), *Revisiting EFL assessment: Critical perspectives* (pp. 133-150). Singapore: Springer.

Kunnan, A. J. (Ed.). (2014). *The companion to language assessment.* Chichester, West Sussex, UK: John Wiley & Sons.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Mizbani, M. & Chalak, A. (2017). Analyzing listening and speaking activities of Iranian EFL textbook Prospect 3 based on Bloom's revised taxonomy. *Advances in Language and Literary Studies,* 8(3), 38-43.

Norris, J. M. (2013). The audio-mirror: Reflecting on student speaking ability. In J. D. Brown (Ed.), *New ways of classroom assessment* (Rev. ed.) (pp. 164-167). Alexandria, VA: TESOL.

O'Sullivan, B. (2014). Assessing speaking. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 156-171). Chichester, West Sussex, UK: John Wiley.

Popham, W. J. (1995). Classroom assessment: What teachers need to know. Needham Heights, Massachusetts: Allyn and Bacon.

Purpura, J. E., & Turner, C. E. (2017). *Learning-oriented assessment in language classrooms: Using assessment to gauge and promote language learning*. Oxford: Taylor & Francis.

Tsagari, D., & Banerjee, J. (Eds.). (2016). *Handbook of second language assessment*. Boston/Berlin: de Gruyter Mouton.

Zafari, S., Heidari Tabrizi, H., & Chalak, A. (2020). The effect of employing critical needs analysis on Iranian intermediate learners' speaking skills improvement. *The Journal of English Language Pedagogy and Practice, 13*(26), pp. 49-74.