

INSIDE THE "BLACK BOX" OF CLASSROOM LANGUAGE TESTS

ELANA SHOHAMY

Tel Aviv University

1. Introduction

Tests are powerful devices as they can determine the future of individuals and programs. Tests are therefore used by those who recognize their power for a variety of purposes, goals and agenda and not necessarily to measure knowledge. Tests are used for creating policies, for introducing new curricula and programs, for defining knowledge, for motivating students to learn and teachers to teach (Shohamy 1994, 1996, 1997). This may be the reason why external tests, those produced by testing agencies and Ministries of Education, get much attention from researchers and test developers. But there are other types of tests, those that are used in the classroom, serve important roles in teaching and learning, yet are being ignored and overlooked. These classroom tests may not have so much power as external tests do as they do not have such significant impact on test takers' lives, but still play an important role in the process of learning. Classroom tests have other traits, other powers, in that they can benefit the instructional process and are extremely useful for teaching and learning. Thus, while external tests are powerful devices that can lead to high stake decisions, tests that are used in the classroom are of lower stakes in terms of their impact on individual lives as they often do not lead to decisions beyond the classroom boundaries. They are nevertheless beneficial from a pedagogical perspective.

The lack of attention to classroom tests is an interesting phenomenon as students are constantly subject to a large number of classroom tests during their schooling while they only rarely take external tests. After all, it is in the classroom where there is a vibrant and dynamic testing culture. It is where tests are used on an on-going basis as part of the learning and teaching process. It

is where quizzes, tests and examinations are used for purposes of providing feedback, assigning grades and conducting evaluations. It is where tests are used by teachers to assess their students, to find out what they have learned to provide them with diagnostic feedback as to the quality and effectiveness of their learning. It is also where teachers obtain information about the effectiveness of their teaching. It is where a variety of assessment tools and devices such as assignments, projects, portfolios, self assessment and other creative procedures are constantly used for multiple of purposes. And it is where tests and quizzes are administered for assessing periodical progress, for obtaining summative information at the completion of teaching units and end of semesters. While a single classroom test cannot determine whether a student will be accepted or rejected to a certain program or whether s/he will enter a specific level of learning, it should also be realized that the cumulative effect of a number of classroom tests can lead to a final score that can eventually lead to a high stake decision as well.

Yet, in spite of these varied purposes and uses of classroom tests throughout the whole education process, only external tests draw the attention of test developers and researchers. Thus, tests such as matriculation examinations in many countries, entrance exams to different programs and graduation tests are the focus of research, theory, development and topics of discussions and debates while classroom tests are being ignored. This holds true for both general testing and language testing. In language testing, for example, it is rare to find any article or piece of research that addresses the issues and uses of language tests in the classroom as most of the energy is devoted to the powerful high stake external tests. The lack of classroom testing practices has led to the situation where theories, procedures and findings from external testing are wrongly applied in language testing in the classroom. Thus, there is hardly any knowledge about classroom tests, how they are used, by whom, for what, when and why.

For example there is no knowledge about:

- how is assessment conducted in the language classroom?
- what type of testing assessment procedures are used by teachers in the classroom?
- do students find these procedures useful?
- do students get meaningful feedback from classroom tests?
- are tests used to measure progress, to grade and/or to discipline and control?
- what goals do classroom tests serve – formative, summative, diagnostic?
- what are the sources from which teachers draw their assessment practices – are assessment practices based on their personal knowledge, training, external tests, or perhaps intuitions?
- how do teachers construct their knowledge about assessment?
- how closely are classroom language tests related to instruction and learning?

- how much background and training do teachers have in testing theory and practice?
- what is the relationship between assessment and learning in the classroom?
- are teachers knowledgeable about the theories and practices of assessment?
- if teachers are familiar with assessment theories, do they find them relevant to their classroom practices?
- how do students view classroom assessment?
- what are teachers' views about language assessment?
- do teachers benefit from the results obtained from classroom tests?

It is, in short, a black box.

2. Background

Within the limited research on classroom assessment it is possible to identify a few topics, those that focus on the type of assessment procedures used (Stiggins – Faires Conklin 1992) and those that focus on the assessment style used by teachers (McMallum et al. 1995); yet, it should be noted that none of these studies focus specifically on language tests. The work of Stiggins and Faires Conklin found that teachers are pretty limited in their repertoire of assessment tools, using mostly tests.

McMallum et al. (1995) examined the effect of the national curriculum on classroom assessment practices and identified three models of teachers. The *intuitive* teacher is characterized as objecting to the imposed system of external national assessment as s/he views it as disruption of intuitive ways of working. Thus, the intuitive teachers are minimal adopters of the national assessment procedures since for them teaching and testing are separate. The *evidence gatherer* is the second type. This model is characterized by having a basic belief in the primacy of teaching, rather than in assessment. Their main method of assessment relies on collecting evidence which they only later evaluate. Assessment is accommodated within the existing system and it is not always planned. Yet, they keep every type of evidence. The third type of a teacher is the *systematic planner* who plans time specifically for assessment which has become part of their practice. For this type the planned assessment of groups and individuals informs future task and design and class work.

This type is also characterized as having a constructive approach to learning by being minimal adopters of the national assessment priorities.

The studies, the results of which will be reported in this chapter, provide some initial insight into this black box. They investigate various aspects of perceptions and practices of assessment in the language classroom.

3. Research

3.1. Study 1

A study by Yariv (1998) addressed issues related to perceptions and practices of classroom tests. It examined a number of questions related to how tests are used in the classroom by English as a Foreign Language (EFL) teachers as well as the perceptions of students and teachers about classroom language tests. This research was concerned with the role that tests play in the composition of the final score, the type of feedback students obtain from classroom tests and whether students felt that classroom tests reflect their actual knowledge. The method of investigation used was a questionnaire administered to EFL 50 students and their 10 teachers.

In terms of the composition of the final score teachers declared that it consists of the following components – 81% of the score was performance on classroom tests; 8% homework; 7% class work; 6% class participation, and 11% other activities. Clearly tests provide the most important input for teachers to arrive at the students' final grade. This finding confirms the research by Stiggins and Faires Conkin (1992) with regard to the important role of tests in classroom learning in comparison to other procedures.

The other question addressed in the study concerned the use of tests as viewed by teachers and students. While teachers claimed that tests are used exclusively to measure language knowledge, students disagreed, claiming that teachers use tests for a variety of other purposes such as creating discipline, controlling and manipulating their behavior. Teachers and students also differed in their views about the kind of feedback they obtain from tests. While teachers claimed they provide students with meaningful and detailed feedback from tests, most students claimed that the type of feedback given to them was neither useful nor helpful. Other disagreements between students and teachers were found in perception regarding the extent to which tests reflect the students actual language knowledge. While teachers claimed to have confidence in their tests as they felt that they accurately reflected students' language knowledge, students thought that tests were not a true reflection of their actual language knowledge. Both teachers and students acknowledged that external tests have a strong effect on classroom tests.

Further information that was obtained from the questionnaires showed that teachers felt they lacked training in testing; yet, those who did obtain training found it to be very irrelevant to the type of issues and problems teachers face in testing in the classroom.

3.2. Study 2

In a number of studies the impact and effect of the introduction of new national external tests on the content of classroom language tests was examined. These

studies (Shohamy 1993; Shohamy – Donitsa-Schmidt – Ferman 1996 and Ferman 1998) used questionnaires, observations and teaching materials for the investigations. They found that high stake tests affect the content and format of classroom tests in various ways. The testing tasks, items and rating scales which were used in the classroom were identical to those used on the external tests. For example, the texts and questions on a classroom reading comprehension tests were the same as those of external tests including texts and multiple choice questions. In an EFL oral test, as well, the type of tasks observed in classroom testing were very similar to those appearing on the external test. It was found that in situations when there is an external test, teachers tend to use tests more frequently in the classroom. In the Shohamy – Donitsa-Schmidt – Ferman (1996) study it was found that when the test is low stake (i.e., no major consequences occur as a result) there is little influence of the tests on classroom assessment practices and more creative type of tasks and activities are used. In all three studies it was found that in situations when teachers are not trained to use tests they turn to the external test as a source of pedagogical knowledge.

3.3. Study 3

In a different study by Ohayon (1998) the goal was to find out about the extent to which there is a connection between two types of classroom tests – bi-weekly quiz and an end-of-year test, conducted over a period of four years and an external examination administered at the end of 12th grade. The comparison focused on content and format of all the classroom tests administered in 9th, 10th, 11th and 12th grades with the external test which is administered at the end of 12th grade.

The research was conducted through a content analysis of all classroom quizzes and end-of-year tests of 10 teachers in comparison to the external tests. The results of the comparison showed that as the time of the final 12th grade external test was approaching, the quizzes and tests administered in the classroom resembled more and more the external test. Specifically, the 9th and 10th grade tests showed only moderate influence of the external tests on those administered in the classroom. These classroom tests included a wide range of tasks – some were more traditional including isolated words and grammar structures, while others were more creative including a variety of communicative and open ended tasks where test takers were required to use language in contextualized ways. Yet, the classroom tests used in the 11th and 12th grades were very similar to the format and content of the external test. Specifically, the results showed that for the traditional teachers the effect of the external test was somewhat positive as it provided them with a testing model which

was better than their own tests. Yet, for those teachers who showed creativity in their 9th and 10th grade tests imitating the external tests in the 11th and 12th grades meant that their tests became more traditional than the ones they used in the 9th and 10th grades. It is clear then, that for some teachers the model of the external test is helpful while for others it is restrictive.

3.4. Study 4

A study by Aazam (1998) examined classroom assessment practices in one school after teachers obtained training in alternative assessment procedures over several months. The researcher used questionnaires and observations to examine the uses of assessment in the classroom. The results showed that there was, in fact, an increased use of alternative procedures in the classroom as the language teachers began to use a variety of alternative procedures such as portfolios, self assessment and peer assessment in their teaching. However, when it came to assigning students the final grade for the semester the teacher counted only the tests and ignored completely the performance of the students on other measures. This provided evidence that while teachers appreciate the use of alternative procedures as part of classroom teaching they did not trust these procedures as valid measures for evaluating students language knowledge in contributing and affecting the final score. Thus, they did not have much trust in such procedures believing that only tests are capable of providing such information.

3.5. Study 5

In the last study reported here (Bahat et al. 1997) the use and application of assessment procedures in a large scale national conducted in Israel in twenty two schools, was studied. The project is an experimental one in which alternative assessment tools are used for arriving at the final grades of students at the end of the 12th grade, instead of the regular matriculation end of high school examination. The 22 schools are exempted from the regular final matriculation examinations, rather their teachers are expected to assign students their final high school grade based on their performance on a large number of alternative procedures. The teachers are being intensively trained by experts to use such procedures and are closely monitored. The research studied one such school. Specifically the question asked regarded the effect of these procedures on classroom practices, on students' and teachers' attitudes and on the definition of language knowledge.

A variety of research methods were used including questionnaires, observations schemes and interviews, administered to teachers, students and the principals. The results showed that while there were some problems in the initial implementation, more varied and creative tasks were used both for teaching

and for testing in comparison to classes that were not included in the study. It was also noticed that assessment was perceived differently, in a more continuous and formative nature. The teachers and students felt they were free to create innovative procedures and original tasks which had not been previously used in the 12th grade and these encouraged independent thinking. Furthermore there was less tension around assessment, the atmosphere was comfortable and students who had obtained low grades before and were left out of the system, now participated more frequently. Students had a growing awareness of assessment and the criteria by which they were being evaluated. They became especially familiar with the rating scales. Probably the most important contribution was that there were expanded definitions of language beyond the definition that was dictated by the national tests which over the years became the defined national language knowledge around the country. Language knowledge was more open and not limited as different materials than these which are normally used came into practice in the classroom. Teachers seemed to be enthusiastic about the new methods and students took a more active role in the evaluation. It became clear that only total removal of tests, at least for a while, could produce a "different" type of assessment behaviors in the classroom.

4. Conclusions

A number of conclusions can be drawn from the above studies regarding classroom assessment. From both the Yariv (1998) and the Aazam (1998) studies it is clear that in spite of the growing popularity of alternative assessment, teachers still count on tests as the major input for final grades, even in situations when teachers are trained in using other procedures and are using them as part of the teaching process. Thus, while teachers may view alternative procedures as helpful instructional tools they do not trust them as tools for providing reliable and valid sources for assigning grades. This finding is similar to the one obtained by Stiggins – Faires Conkin in 1992. Yet, from the Bahat et al. (1997) study it can be learned that the use and application of alternative procedures are successful when tests are totally removed from the schools' agenda. It is clear that there is still a long way before teachers will view alternative procedures as a true alternative and will trust them the way they trust tests. This may derive from the lack of training on how to incorporate alternative assessment in the classroom, or from the demand of administrators for grades that can be defended or from the fact that teachers are still hesitant in using procedures that are viewed as "subjective". Yet, it may also derive from the fact that there are no acceptable methods to examine the qualities of such devices.

This may be related to the conclusions from the Yariv (1998) study indicating that teachers feel they lack the knowledge and training required to practice assessment procedures; worse still is the perception of those who did obtain

training that it is not very relevant and very useful for their classroom teaching. Interesting is also the finding that teachers and students have different views regarding the roles, values and uses of tests.

In terms of the type of tests used in the classroom, the studies by Shohamy (1993), Shohamy et al. (1996) and Ferman (1998) indicated that external tests have a very strong impact on the content and format of classroom tests. In fact teachers imitate entirely the content and format of external tests. Thus, external tests provide *the* model for classroom tests, that is imitated and replicated in the classroom. Thus, the external test is not only powerful in leading to high stake decisions but also in providing teachers with the pedagogical knowledge from which they base their own knowledge of how classroom tests should be like. Instead of developing their own classroom assessment models teachers tend to imitate and replicate the external model, probably believing that if it has the power, it must be better and right. Shohamy (1993) showed that in situations when teachers lack knowledge of methods of testing they tend to turn to the external test as the main source of pedagogical knowledge. Whether this is a result of the power and status of external tests, the lack of training, or the irrelevance of training is an open question that needs to be examined. Yet, as was learned from the Ohayon (1998) study, having the model of the external test offers benefits for teachers who lack knowledge in assessment, if these external tests are well constructed and reflect current practices of assessment. Yet, for the more highly qualified teachers who are familiar with current ways of assessment the external test may narrow and restrict their creativity.

5. Implications

The results from the above studies provide initial information about patterns of classroom assessment. In terms of the questions posed at the beginning of this paper, these results are just a drop in the bucket in the area of classroom assessment. There is a need to study this assessment context and to come up with theories and hypotheses about the use of assessment in the classroom. It is clear that although classroom assessment differs from external assessment in terms of goals, purposes and context external assessment is the dominant paradigm in many classrooms. Yet, since the goals of classroom assessment are different than those of external assessment, it is clear that there is a need for a different theory for classroom assessment. When theories like this are established, other strategies may be used. It is likely that classroom assessment requires different types of assessment procedures as their uses and purposes differ. Furthermore, it is very likely that different types of quality control are needed as well. Thus, the types of validity and reliability which are practiced for external tests may not be applicable for classroom assessment. It is obvious that many of the assumptions regarding reliability cannot even be applied in

the classroom. For example, item independence may not be possible to achieve in the classroom, test retest is problematic as well, and internal consistency is not feasible to carry out on classroom tests which are generally very short. Similarly, in terms of items; multiple choice and other objective items were developed mostly for external assessment so they can provide items that can be scored fast, often with the use of a machine. Yet, these types of items became very popular and common practices in many classrooms. They are clearly inappropriate for this context. What are, then, the type of items and tasks most suitable for the classroom assessment? What are the ways of using assessment in the classroom? How should grades be assigned, if at all? If classroom assessment involves continuous feedback, what procedures are most appropriate for that purpose?

Since classroom tests are not independent but rather connected to instruction and learning it is most likely that assessment needs to include these features. Thus, it is very possible that theories of classroom assessment need to address the unique features of assessment as part of learning – it therefore needs to focus on making them more interpretive, include formal and informal components, be more diagnostic and based on a dialogues and mutual construction by teacher and students. These unique features imply that there is a need to develop a special framework and theories to inquire various aspects related to classroom assessment. Research and development in this area of classroom assessment, involving all those who are part of the classroom context – teachers, students and possibly parents, is needed as classroom assessment plays such a major role in learning and acquiring language knowledge.

Finally, an interesting topic of research by itself is to examine the politics and ideology of research on external tests and to find out the reasons why external tests get all the attention while classroom tests do not. This question is related to the politics of research and is similar to an attempt to find out why in the field of medicine certain drugs get researched and manufactured while others do not. Are these driven by economic and market forces, or by the status and power of the object of research? Is it a result of the power groups standing behind it, or because groups use tests to promote certain agendas taking into account the power they possess. Stiggins and Faires Conklin (1992: viii) claim that “Due to a lack of classroom-level research, we fail to understand the assessment tasks demands faced by teachers. The apparent extreme complexity of classroom assessment environments and issues has served as a barrier that has kept researchers from conducting the needed research.” They claim therefore that it is time to confront this complexity.

REFERENCES

- Aazam, Umaima
 1988 Shimus be'haaracha chalufit b'aarbaa batei seifer (The use of alternative procedures in four schools.) [Unpublished M.A. thesis, Tel Aviv University.]
- Alatis, James E. (ed.)
 1994 *Georgetown University Round Table on Language and Linguistics*. Washington, D.C.: Georgetown University Press.
- Bahat, Susan – Judy Kemp – Ella Shynin – Samya Kara – Nahaya Natur – Sharon Reis
 1997 Haaracha Halufit: Case Study shel ehad mibatei haseifer shmishtatfim beproyekt, bagrut 2000. (Alternative assessment: Case study of one school participating in Bagrut 2000.) [Seminar paper in research on language testing, Tel Aviv University.]
- Ferman, Irit
 1998 The impact of a new English foreign language oral matriculation test on the educational system. [Unpublished M.A. thesis, Tel Aviv University.]
- McCallum, Bet – Caroline Gipps – Shelley McAlister – Margaret Brown
 1995 "National curriculum assessment: Emerging models of teacher assessment in the classroom", in: Harry Torrance (ed.) 57-87.
- Ohayon, Rachel
 1998 Hakesher bein Mivchanei Kita Vemivchanei bagrut (The connection between classroom tests and the Bagrut.) [Unpublished M.A. thesis, Tel Aviv University.]
- Shohamy, Elana
 1993 "*The power of tests*": *The impact of language tests on teaching and learning*. (NFLC Occasional Papers.) Washington, D.C.: National Foreign Language Center.
 1994 "The use of language tests for power and control", in: James E. Alatis (ed.), 57-72.
 1997 "Critical language testing and beyond", (Plenary address at the annual American Association for Applied Linguistics Conference, Orlando, FL, March 1997.)
- Shohamy, Elana – Smadar Donitza-Schmidt – Irit Ferman
 1996 "Test impact revisited: Washback effect over time", *Language Testing* 13, 3: 298-317.
- Stiggins, Richard – Nancy Faires Conkin
 1992 *In teachers' hands*. Albany, NY: State University of New York Press.
- Torrance, Harry (ed.)
 1995 *Evaluating authentic assessment*. Buckingham – Philadelphia: Open University Press.
- Yariv, Ziva
 1998 Shimush Bemivchanei kita Behoraat Anglit bezika Lebhinot Habagrut. (The use of classroom test in teaching English in the upper level in relation to the matriculation examination.) [Unpublished M.A. thesis, Tel Aviv University.]