The Language Assessment Process: A "Multiplism" Perspective

Elana Shohamy and Ofra Inbar, Tel Aviv University

Assessment refers to the processes through which judgments about a learner's skills and knowledge are made (Bachman, 1990, Lynch, 1996; McNamara, 1996). The word *assessment i*s derived from the Latin *assidere*, which means "to sit beside", thus allowing the bystander to observe the learners and gather information.

What is assessment used for? Shepard (2000) divides assessment purposes into three major categories: **administrative**, **instructional**, and **research**. The purposes he outlines in the "ad- ministrative" category include general assessment, placement, certification and promotion. The "instructional" category includes the use of assessment for diagnosis, for evidence of progress, for providing feedback to the respondents and for evaluating the curriculum. The purpose of as- sessment in the "research" category entails research experimentation, knowledge about language learning and knowledge about language use.

The process of conducting assessment includes a series of phases (see below), regardless of the specific assessment instrument (or procedure) that is being employed. It begins by setting a pur- pose, defining the relevant language knowledge to be assessed and selecting a suitable assess- ment procedure from various available alternatives. Once purpose, language knowledge, and assessment procedures have been decided, the actual assessment task (or tasks) and items are designed and produced. When the assessment instrument is "ready", it is administered to the language learners. Subsequent steps in the process are to assess the quality of the instrument itself, to examine validity and reliability, and to note any difficulties which may occur before, during or after the administration. The assessor will then proceed to interpret the results and finally report them to the various parties (i.e. the stakeholders) involved.

The Language Assessment Process

- Determining the purpose of assessment
- Defining language knowledge to be assessed
- Selecting the assessment procedure
- Designing items and tasks
- Administering the assessment tool(s)
- Determining the quality of the language sample/answers produced
- Assessing the quality of the procedures
- Interpreting the results
- Reporting the results

Deciding which *assessment tool(s)* to use depends on the purpose of assessment and on how lan- guage knowledge is defined. Shohamy (1998) elaborates on these points suggesting the use of a "multiplism" approach to language assessment, whereby multiple options are available at each phase of the assessment process.

The Multiplism Concept in Assessment

When planning and creating a language assessment instrument we need to consider many dif- ferent variables: we need to think about how well the specific instrument we chose represents the topics it aims to assess, what it looks like, its fairness and ethicality in terms of the students, the suitability of the type of item chosen and the feedback it provides for on-going teaching and learning.

Three pertinent questions need to be considered before we start designing the instrument:

1. What is the purpose for conducting this assessment procedure? For example: Will the as- sessment instrument be used for checking learner achievement of what has been taught? Is it intended to uncover what students know in order to assign them to groups or levels or place them into a given program? Will it be used for reporting progress to external agencies or for providing research data?

2. H**ow is the language knowledge to be assessed defined?** Each of the above purposes requires a different focus thus necessitating different definitions of the language knowledge base to be as- sessed. The teacher designing an achievement test, for example, might define such knowledge as the content of the unit taught, the theme, relevant functions, lexical items and grammatical struc- tures. The targeted knowledge of a test given at the workplace to predict whether to accept or reject potential candidates will be directly related to the specific job description required. Hence, if a translating agency is seeking interpreters who are proficient in certain languages and also knowledgeable in specific domains (like commerce), the defined and evaluated knowledge base will include familiarity with the domain in the specified language(s) and ability to transfer that knowledge from one language to another. On the other hand, a university entrance exam which aims to assess academic performance will delineate the language knowledge required of students in institutions of higher learning, such as reading academic texts, analyzing and synthesizing data and writing position papers. Thus the purpose for assessment and the targeted language knowledge being assessed are inseparable.

The assessment procedure is therefore considered valid if the testing instruments actually mea- sure the knowledge it sets out to measure and provides the users with the data they were seeking. If we take for example an achievement test created by a classroom teacher to evaluate whether or not students have acquired particular knowledge, that test will be valid if it does indeed make the information about these

specific learning outcomes available. In terms of the workplace, and the translation test cited above, the employers should be able to decide who to hire for the translation task based on the outcomes of the assessment procedure they have developed and implemented.

3. What instruments or assessment procedures will be chosen to elicit the required language knowledge? Unless a suitable tool is developed learners will not be able to demonstrate that they have acquired the targeted language knowledge for the stated purpose. Consider the following situation: A teacher wants to check interactive spoken ability. The defined knowledge includes using social language skills such as greetings, ability to request and provide information, using appropriate language register, etc. If the chosen assessment procedure is a written dialogue the relevant information as to the learner's ability to interact socially will probably not be obtained.

CALPER Professional Development Document 3

Choice of a spoken simulated interview format, on the other hand (rather than the written dia- logue), will allow the test takers to demonstrate their interactive ability (or lack of it) in a far better manner.

There is disagreement and great variability among educators and researchers as to what actually constitutes language knowledge as well as to the suitable procedures for assessing this knowledge. A survey of these different opinions shows that they stem from, and correlate with current language teaching approaches as specified by theories of language learning. The following section elaborates on how language knowledge was perceived in different time periods and the impact these percep- tions had on shaping language tests.

Teaching Approaches, Language Knowledge and Testing Methods

Tests in the *discrete-point teaching era* reflected the view that language knowledge is comprised of isolated items (Spolsky, 1975). Thus the testing of specific language structures or decon- texualized vocabulary items via objective closed test items constituted the dominant assessment format during this time. In the period when language was seen from a more global integrative perspective and testing became more contextualized, relating to full texts rather than discrete components of language, integrative methods of assessment such as the 'cloze' procedure were widely used. *Communicative language teaching* emphasized language use for real direct purposes. According to Canale and Swain (1980) (following Hymes, 1974), communicative competence was seen to consist of linguistic competence, sociolinguistic competence, discourse competence and strategic competence. In order to measure these notions of competence, testing procedures were expected to simulate functional and relevant language use as authentically as possible. *Performance teaching* has added to the previous perspective the relevance of the interaction among language knowledge and specific content areas and contexts. Subsequently matching language assessment instruments suited for particular situations and audiences were designed.

Thus, just as the discrete-point approach to language knowledge created in turn tests of specific disconnected language items, so the current perception of language as a complex system has impacted the latest view of language testing. In this case targeted language knowledge refers to what language users can *do* with the

language in authentic situations and to their ability to understand and produce language samples appropriate for particular contexts, rather than to merely recognize specific language components (Fulcher, 2000). Developing such linguistic competence calls for the integra- tion of **tasks** that simulate real language use, and involve the learners in a variety of oral and written interactions with speakers of the targeted language. In order to fully represent the students' ability the assessment data needs to sample an array of domains of language use. Hence both productive (writing and speaking) and receptive (reading and listening) abilities ought to be assessed as well as the ability to integrate these in a way that characterizes authentic language behavior: when we talk to someone we both listen and speak and sometimes refer to written notes, or read a text to make a point. There are also different objectives within each skill depending on the test purpose as we described above. In order to qualify as a capable listener in the target language, for example, the student will be assessed on abilities to comprehend a lecture, radio talk shows and recorded phone messages. As Bachman and Palmer (1996) claim:

> [...] it is not useful to think in terms of 'skills'. But to think in terms of specific activities or tasks in which language is used purposefully. Thus, rather than attempting to define 'speaking' as an abstract skill, we believe it is more useful to identify a specific language use task that involves the activity of speaking, and describe it in terms of its task character- istics and the areas of language ability it engages. (p.76)

Since language knowledge consists of numerous variables, a single testing procedure cannot adequately assess them all, and drawing conclusions as to the individual's knowledge on the basis of a single tool is problematic. This creates the need to develop multiple procedures for collecting data for various purposes. **Language assessment tools** will then include, for example,

- projects
- putting on a play
- creating a restaurant menu
- simulating "real life" situations (e.g. purchasing goods at a store)
- reporting an event
- creating a game or video-clip
- corresponding in writing for various purposes.

In addition, learners need to be able to ascertain their own abilities so that they can find ways to improve them. They will therefore be engaged in self-assessment as well as in the assessment of their peers.

Assessment is thus viewed as an **on-going process** bound up with the learning process rather than as a single episode that occurs at the end of a teaching unit. Classroom teachers are encour- aged to use a variety of assessment tools, both formal (like tests), or informal (like observations). Classroom assessment focuses on both the **process** and the **product** components of language use. In teaching and assessing reading and listening, for example, the process relates to the strategies used to access a written or oral text, while the product is actual comprehension of the text.

Assessing language abilities through employing "portfolios"embodies what we discussed so far, for it includes different representations of a learner's language knowledge and ability to perform different tasks. A **portfolio** is defined by SABES (System for Adult Basic Education Support) as:

PORTFOLIO DEFINITION

a collection of work, usually drawn from students' classroom work. A portfolio becomes a portfolio assessment when (1) the assessment purpose is defined; (2) criteria or methods are for determining what is put into the / whom, and when; and (3) criteria for ther the collection or individual pieces are identified and used to make about performance. Portfolios can be assess student progress, effort, and/or t, and en- courage students to reflect ning

URL: http://www.sabes.org/assessment/glossary.htm

Each piece of work in the portfolio (e.g. reports, projects, self or peer assessment, etc.) allows the language teacher to elicit different language samples and to gain added knowledge about differ- ent facets of the learner's language ability. Once all of these pieces are incorporated, a more com- plete "picture" of the learner's capabilities will emerge. This allows the teacher to better relate to particular needs, and provide focused and efficient feedback to the student. The student in turn is an active participant in both choosing the language samples s/he is judged by and self-assess- ing them along with others. It is this concept of "**multiplism"** (from Cook, 1985) which Shohamy (1998) thus proposes to apply to current perspectives in language testing.

The notion of multiplism in language assessment takes a broad view of language knowledge and assessment. It refers to multiplicity in a number of areas:

> [...] multiple purposes of language assessment, multiple definitions of language knowl- edge, multiple procedures for measuring that knowledge, multiple criteria for determin- ing what good language is, and multiple ways of interpreting and reporting assessment results. (Shohamy, 1998, p. 242).

It includes both **formative** (on-going) and **summative evaluation** (at the end of a process), **achievement** (assessing what was learnt in a particular program) as well as **proficiency** (general language capacity unrelated to a particular language program) assessed via a a wide array of assessment procedures. The multiple approach can be implemented in various phases of the as- sessment process and relates, among other things, to the pertinent issues discussed above: setting the purpose for assessment, defining the language knowledge and outcomes, and determining what assessment instruments will be used in each case.

1) Multiple purposes of assessment. Here multiplism refers to the different reasons one may have for

using assessment, such as checking achievements and progress, predicting success, mo- tivating, categorizing and exercising power.

Multiple Purposes of Assessment

- •Predicting success
- •Placing students according to proficiency levels
- •Accepting or rejecting students to a language program
- •Providing feedback on students' learning
- •Following the progress of individuals and groups
- •Motivating students to learn the language
- •Disciplining learners
- •Exercizing power in the language classroom
- •Conducting research on various facets of language study

In terms of defining language knowledge and outcomes, these will depend on the set purpose for assessment. Models of language knowledge, however, have attempted to provide general frame- works for language knowledge, such as the Canale and Swain 1980 model mentioned above and the Bachman 1990 model of communicative competence. Language ability is also defined in terms of the tasks learners can perform in a language. Having performed the tasks to varying degrees of success learners are accordingly classified as novices, intermediate or advanced users of the language.

2) Multiple assessment procedures. While in the past 'tests' were the predominant assessment format used, multiple assessment procedures are currently employed. These refer to the range of assessment options from open informal instruments such as unstructured observations to per- formance tasks of various sorts which simulate authentic language performance for a variety of purposes. Self and peer assessment procedures have also become part of the language assessment repertoire, used either to supplement other measures or on their own as stand-alone procedures. Each of these is chosen on the basis of its characteristic features and suitability for the testing situation (for example, costs and availability of trained raters).

It is important to note that although tests are no longer the only means for carrying out an assess- ment they are still recognized as valid and valuable instruments for particular purposes, such as certain forms of summative assessment or external assessment used for classification. Norris (2000) mentions four dimensions which determine test use:

- who uses the test;
- what information should the test provide;

- why, or for what purpose, is the test being used, and
- what consequences should the test have.

It is up to the test writer to decide what kind of test will be designed on the basis of these four dimensions. The following table lists some of the many procedures a teacher/examiner can choose from.

Some of the Multiple Methods of Assessment

Portfolios Homework Self-assessment Oral debates Tests Dramatic performances Projects Role plays Simulations

Learning logs Interviews Peer-/Group-assessment Check lists Diaries Observations Presentations Dialogue journals Rubrics

3) Multiplism in designing items and tasks. A wide variety of both items and tasks are available for constructing assessment procedures. The term "**item types"** often refers to techniques for assessing mostly the comprehension skills (reading and listening) and includes procedures such as matching, true/false, multiple choice, cloze passages and open-ended questions. **"Tasks"** are

used more often for examining the production of oral and written language samples, and include formats such as interviews or essay writing. This division between productione and comprehen- sion skills is not always applicable for many tasks, especially the more complex ones such as projects and presentations, which require the integration of different language skills and language functions. In order to carry out a project, for example, a learner is required to summarize the main points from different sources (comprehension) and then react to the ideas found and create new ones (production). Choosing which tasks or items to use depends once again on their relative merits and degree of suitability to the assessment purpose and context. Some of the commonly used items and tasks are listed in the following table:

Multiple Ways of Designing Items and Tasks

Multiple Choice True / False Open-ended Questions Essay Questions Summaries Cloze Passages Tasks Role Plays Reporting

4) Multiple ways of administering. Rather than the traditional single administration of a pa- per and pencil test, present assessment administration conditions vary to include on-line testing, video and audio

components as well as individual and small group assessment formats often via computers. The testers may be the teachers or external assessors and administration may be done overtime as a formal or informal procedure. Examples of various administration forms are:

Multiple Ways of Administering Assessment

- **one-to-one administration**
- **paper and pencil format**
- **audio-taped tests**
- **visual stimuli and questions**
- **computer-administered assessment**
- **in-classroom vs. take-home**
- **on site assessment (at the workplace)**
- **formal and informal administration**

5) Multiple criteria for determining language quality. Determining criteria for assessment will evolve from the test purpose, the type of language knowledge and ability targeted and the tasks or items chosen. The response may be one-dimensional as in closed item formats (e.g. multiple choice or matching item types), or open to multiple interpretations as in a performance task. In the

first case scores will be added up numerically. In the latter case scoring criteria will be determined and presented in the format of rubrics, which incorporate task relevant dimensions presented in hierarchical descriptors. Criteria may also appear in the form of rating scales, either holistic scales (rating scales which assess global language ability) or analytic scales (rating scales which focus on a specific language component such as fluency or accuracy). The actual assessment criteria may be determined according to given standards or guidelines such as the ACTFL Guidelines. The following are some of the different criteria for judging language ability.

Multiple Criteria for Determining Language Quality

total score standards, benchmarks, competencies, can-dos,
bandscales diagnostic criteria holistic rating scales analytic
rating scales rubrics guidelines (e.g. ACTFL, ISLPR) native /

6) Multiple criteria for determining the quality of assessment procedures. Assessing the qual- ity of assessment procedures involves examining the reliability and validity of the tools used. Validity comes from the word *valid*, i.e. *has value*. An assessment tool is perceived as being valid if it actually assesses the language abilities it aims to assess. In classroom teaching this would mean that the instrument matches the objectives set by the teacher/assessor which were formulated based on, and in accordance with, the teaching that occurred prior to the assessment activity. We distinguish among a number of validity types, each relating to a different aspect of the assess- ment: **content**, **concurrent**, **predictive**, **construct** and **face validity**. **Content validity** is the most relevant validity for the classroom teacher, since it examines the extent to which the assessment measure, task or test, represents the content to be assessed. In terms of advanced language abil- ity, for instance, this means that the assessment tool represents the specifications described in the curriculum standards. The higher the coordination between the tool and the standards or aspects it intends to assess the higher the content validity the tool has. **Concurrent validity** examines whether a particular assessment tool yields similar information as another tool intended to assess the same knowledge. **Predictive validity** examines if the test can correctly predict success in a given language function or context. In other words, whether a testee who succeeded in obtaining a high score on an English for Academic Purposes test will actually perform well in this area in the future, i.e., manage to read academic texts as required. **Construct validity** examines whether the assessment tool is in line with the current theory of the trait being examined. A listening test, for example, will have high construct validity if it reflects current theories of comprehension processing in terms of meaning construction. **Face validity** examines whether there is a match between what he test actually look like and what it is supposed to test (more on this issue in the section on the testing process).

In recent years, notions of construct validity have been substantially expanded to include issues related to the consequences of tests, specifically to the social and educational impact that tests have on test takers and on learning. Messik (1989), who was the first to introduce this notion, presents an expanded view of the responsibility of testers to include the consequences of tests as an integral part of construct validity. This implies a need to examine how the tests are actu- ally used and if there is evidence as to their positive impact and sound values (Kunnan, 2005; McNamara, 2001, Shohamy, 2001). Whether these are separate types of validity or an integral part of construct validity is a point of debate (Popham, 1997; Shepard, 1997).

Reliability refers to the extent to which the test is consistent in its score, thus indicating whether the scores are accurate and can be relied upon. This concept takes into account the error which may occur in the assessment process. Just as with other forms of measurements, such as scales de- signed to measure weight or temperature, some errors may occur in the process. The score is seen to consist of the true score and a measurement error and together they constitute the observed score which the student receives. The source for measurement error varies: it may stem from the raters' subjective assessment, from the

difference between assessment measures designed to test the same subject area, from external conditions which affect scores such as technical facilities, and how the items on the test relate to one another. The standard error of measurement (SEM) is an estimate of the error and serves to interpret individual test score within probable limits or intervals. Thus, if an observed score is 70 and the SEM is 3, the student's true score will fall within the range of 67 to 73. Obviously, the smaler the SEM, the more reliable the test will be, because the observed score will be closer to the true score.

Reliability measures help us estimate the error in the score: the higher the reliability measure the lower the error and the more reliable the score. Some assessment measures are viewed as more reliable since the possibility of error is limited: for example when scoring closed-item formats (such as multiple choice or true false) where there is a predetermined single answer there is less of a chance that ratings will be influenced by personal subjective variables than in open-ended tasks, where the answers vary and the raters have to use different criteria to determine the score.

Agreement among raters is referred to as **inter-rater reliability**. Sometimes the same rater may assign different assessment scores or evaluations due to a variety of reasons (physical condi- tions, fatigue, effect of previous grading of assignments etc.). In this case there is a problem with **intra-rater reliability**. Both types of rater reliability are important for items and tasks of an open nature (for instance written compositions and oral interviews) where it is likely that there will be disagreements with regard to the quality of the language sample.

Other reliability measures are **test-retest reliability** (the extent to which the test scores are stable from one administration to the next) and **internal consistency** (the extent to which the test items measure the same trait).

A test may be reliable (consistent score) but not valid, i.e. the score is reliable but the contents of the test do not reflect the test writer's objectives or what students have learned. In order to determine the quality of test items, analysis of the levels of difficulty of each item is examined, i.e., how many of the test takers got the item correct, and the discrimination index per item will be calculated, i.e., does the item discriminate between weaker and stronger learners. These indices are especially important when using instruments whose purpose is to select learners according to proficiency levels. To summarize, the type of criteria used for determining the quality of the assessment procedures can be:

- different types of item analyses (difficulty, discrimination, etc.)
- different types of reliability and validity

7) Multiple ways of interpreting and reporting results. The interpretation of outcomes of as- sessment as satisfactory or not depends on the particular situation, on the purpose for which the assessment is given and on learner-related variables. If the person being tested is a new im- migrant, for example, interpretation of the assessment results will need to take into consideration the length of stay in the target

language speaking environment, the kind of language program s/he is enrolled in and the willingness of the learner to invest in learning the language.

In addition to the testees, results can be reported to various other stakeholders -- including par- ents, bureaucrats, employers, institutions. The manner of reporting will change depending on its purpose and future use: if the assessment procedure was conducted to motivate and or monitor on-going progress the results will be discussed with the learner in detail and feedback provided. There are multiple users and stakeholders who are interested in and impacted by the reported results (Rea-Dickens, 1997) and the reporting format will differ depending on the relevant par- ties, on whether the report is intended for the students, parents, teachers and/or other academic or administrative stakeholders. Results can be reported in the form of a dialogue between the as- sessor and the person assessed, or a conference which would include other relevant participants in addition to the two mentioned, for example other teachers who teach the same individual, a counselor, or the student's parents.

The multiple means of conducing these phases in the assessment process are summarized for each phase below:

Multiple Ways of Interpreting Results

- Context-embedded interpretation
- Dialoguing with the student (over email, for example
- Holding a assessment conference (with the student and/or other participants)

Multiple Ways of Reporting Results

- as test scores
- providing diagnostic information
- notifying as Pass or Fail
- comparison of grades (to other populations)
- creating learner profiles
- providing verbal descriptions and interpretive summaries
- reporting in form of narratives
- creating progress reports

Now that we have reviewed the concept of 'multiplism' in the assessment process let's look at an example which demonstrates this notion.

Jack Fillmore has studied Japanese for 6 years and is in an advanced language learning class. He is also studying social studies in Japanese and is now concluding the first se- mester of the final year of his studies. Throughout the semester Jack was assessed with a variety of tools in both his

Japanese language class and his social science classes. The tools included were: tests, written and oral performance-based tasks (projects, a written and oral report, a book task, simulated conversations with various interlocutors). Jack has chosen to include some of the tasks in a portfolio. The portfolio contents were chosen according to a list of required and optional components provided by the teacher. The portfolio was handed in to the teacher and a grade was assigned according to given criteria. Jack has also self-assessed the portfolio according to the same criteria (both the different components and the portfolio as a whole). Following the assessment Jack and two of his instructors – the teacher of Japanese and one of his content course teachers – conduct an assessment conferencing session. The participants, including Jack himself, discuss the achievements in the various areas, exchange views on certain portfolio components and their quality, and provide feedback on what needs to be improved. In this conference the teachers and the student map Jack's needs in view of the evidence presented. The comprehensive picture they get from the multiple sources allows them to do so fully by relating to both Jack's overall ability as well as to specific language components. At the end of the conference the participants draw a profile of Jack's language abilities and needs. This will serve to plan future work and required progress for both the teachers and the student. A report summarizing the conference decisions will be sent to Jack's parents and to the school ad- ministration.

The notion of multiplicity is thus exercised in the above example in a number of ways:

- Use of multiple assessment tools
- Including a number of assessors
- Multiple criteria for determining language ability
- Multiple ways of administering assessment
- Multiple ways of reporting assessment data
- Multiple stake holders

In this *CALPER Professional Development Document* we have attempted to demonstrate that al- though the language assessment process follows a set format of clearly defined phases, there are different possibilities to choose from at each phase. We have traced these different phases showing the multiple ways for conducting each of the steps along the way. The choice of which option to use will depend on the purpose of the specific assessment being conducted, the defini- tion of the language being assessed and the instruments or procedures used to elicit the language knowledge.

Finally, it is important to note that throughout the assessment process the assessor needs to con- sider the ethical and moral questions as well as dilemmas involved in designing and administer- ing the assessment instrument. These can influence the decision as to whether to administer the tests, and include issues such as possible biases against certain groups in the population and the decisions that will be made on the basis of the results: What will the consequences of these decisions be that are based on the assessment? Will certain segments of the population be af- fected more than others? Will the scores provide justification for denying or granting rights and privileges to certain sectors? Will the administration of the tests affect the status of the language in a given context, highlight one language and down grade another?

Thus the assessment process focuses not only on the language and assessment methods but also on wider social concerns. These need to be constantly attended to since the administration of the assessment procedure may lead to unwanted consequences in terms of educational as well as societal and moral issues.

References:

ACTFL Proficiency Guidelines. American Council for the Teaching of Foreign Languages http://www.sil.org/lingualinks/LANGUAGELEARNING/OtherResources/ACTFLProficiencyGuidelines/contents.htm Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press. Bachman, L.F., & Palmer, A. S. (1996).). *Language testing in practice.* (ch. 2: Test Usefulness: Qualities of language tests). Oxford: Oxford University Press. Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second-language teaching and testing. *Applied Linguistics, 1,* 1-47. Cook, T. (1985). Postpositivist critical multiplism. In R.L. Shotland & M.M. Mark (Eds.), *Social science and social policy* (pp. 21-62). Beverly Hills, CA: Sage. Fulcher, G. (2000). The "communicative" legacy in language testing. *System, 28*, 483-497 Hymes, D. (1974). *Foundations in sociolinguistics.* Philadelphia: University of Pennsylvania Press International Second Language Proficiency Ratings (ISLPR). URL: http://www.gu.edu.au/centre/call/content4.html Kunnan, A. J. (2005). Language assessment from a wider context. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 779-794). Mahwah, NJ: Lawrence Erlbaum. Lynch, B. (1996). *Language assessment and program evaluation*. Cambridge: Cambridge University Press. McNamara, T. F. (1996). *Measuring second language performance*. London: Longman McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press. Messick, S. (1989). Validity. In R.L. (Ed.), *Educational Measurement* (3rd edition.) (p. 3-104). New York: American Council on Education Norris, J. (2000) Purposeful language assessment: Selecting the right alternative test. *English Teaching Forum, 39* (1). Available online at URL: http://exchanges.state.gov/forum/vols/vol38/no1/p18.htm Popham, W.J. (1997). Consequential validity: Right concern - wrong concept. *Educational Measurement: Issues and Practice*, *16,* 9-13 Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16,* 5-8, 13 Shepard, L. (2000) The role of assessment in a learning culture . *Educational Researcher*, *29*, 1-14.

Available online at URL: http://35.8.171.42/aera/pubs/er/pdf/vol29_07/AERA290702. pdf Shohamy, E.

(1998). Evaluation of learning outcomes in second language acquisition: A multi- plism perspective. In H. Byrnes (Ed.), *Learning foreign and second languages* (pp.238-261). New York: The Modern Language Association of America. Shohamy, E. (2001). *The power of tests*. Harlow, England: Pearson Education. Spolsky, B. (1975). Language testing – the problem of validation. In L. Palmer & B. Spolsky (Eds.), *Papers on language testing 1967-1974* (pp. 147-153). Washington, D.C.: TESOL.

Please cite as:

Shohamy, E., & Inbar, O. (2006). "The language assessment process: A "multiplism" perspective", (CALPER Professional Development Document 0603). University Park, PA: The Pennsylvania State University, Center for Advanced Language Proficiency Education and Research.

This CALPER Professional Development Document was developed and produced with funds from a grant awarded to CALPER by the United States Department of Education (CFDA 84.229, P229A020010). However, the contents do not necessarily represent the policy of the Department of Education, and one should not assume endorsement by the Federal Government.

Center for Advanced Language Proficiency Education and Research The
Pennsylvania State University 5 Sparks Building University Park, PA
16802