Center for Advanced Language Proficiency Education and Research (CALPER)

The Language Assessment Process: A "Multiplism" Perspective

Elana Shohamy & Ofra Inbar

CALPER Professional Development Document 2006

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* is 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, research**. The purposes outlined in the administrative 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 the evaluation of the curriculum. The third category refers to research experimentation, knowledge about language learning and knowledge about language use.

The process of conducting assessment includes set phases (see below), regardless of the assessment instrument (or procedure) being employed. It begins by setting a purpose, defining the relevant language knowledge to be assessed and selecting a suitable assessment procedure from the various available alternatives. Once the purpose, language knowledge, and assessment procedures have been set, the actual task and items will be designed and produced. When the assessment instrument is ready, it will be administered to the language learners. The next step will be to assess the quality of the instrument, examining validity and reliability, and noting any difficulties which may have occurred before, during or following the administration. The assessor will then proceed to interpret and make sense of the results and finally report them to the various parties (i.e. 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 for assessment and on how language 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 approach is outlined in detail in the following sections.

The Multiplism Concept in Assessment

When planning and creating a language assessment instruments we need to consider many different variables: we need to think about how well this tool 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 assessment instrument be used for checking achievements of what was taught? Is it intended to see 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. How 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 assessed. The teacher designing an achievement test, for example, might define the knowledge as the content of the unit taught, the theme, relevant functions, lexical items and grammatical structures. The targeted knowledge of a test given at the workplace to predicts 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 define 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 measure 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 the classroom teacher to evaluate 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 used.

3. What instruments or assessment procedures will be chosen to elicit the required language knowledge? Unless a suitable tool is developed the 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 comprises of 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 in this particular area will probably not be obtained. Choice of a spoken simulated interview format, on the other hand (rather than the written dialogue), will allow the test takers to demonstrate their ability (or lack of it) in a far better manner.

There is evident disagreement and great variability with regard to what actually constitutes language knowledge as well as the suitable procedures for assessing this knowledge. A survey of these different opinions shows that they stem from and correlate with reigning language teaching approaches in particular periods in terms of theories of language learning. The following section elaborates on how language knowledge was perceived in different periods and the impact these perceptions had on shaping language tests.

Teaching Approaches, Language Knowledge and Testing Methods

Tests in the *discrete point teaching era* reflected the view taken of language knowledge as comprising of isolated items (Spolsky, 1975). Thus the testing of specific language structures or decontexualized vocabulary items via objective closed test items, constituted the overriding assessment format during that era. In the period when language was perceived as a more global *integrative perspective* and testing became more contextualized, relating to full texts rather than discrete language, using integrative methods of assessment such as the 'cloze' procedure. *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 has the current perception of language as a complex system 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 integration 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 activity characteristics and the area 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 for the

development and use of multiple procedures for collecting date 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, the learners need to be able to determine 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 a single episode that occurs at the end of a teaching unit. Classroom teachers are encouraged 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.

Assessing language abilities through employing "portfolios" embodies the characteristics of these notions for it includes different representations of a language learner's language knowledge and ability to perform different tasks. A portfolio is defined by SABES (System for Adult Basic Education Support) as:

> 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 different facets in the learner's language ability. Once all of these pieces are incorporated, a more complete "picture" of the learner's capabilities will merge. This allows the teacher to better relate too 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-assessing 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.
>
> 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 made clear for determining what is put into the portfolio, by whom, and when; and (3) criteria for assessing either the collection or individual pieces of work are identified and used to make judgments about performance. Portfolios can be designed to assess student progress, effort, and/or achievement, and encourage students to reflect on their learning
>
> URL: http://www.sabes.org/assessment/glossary.htm

The notion of 'multiplism' in language assessment therefore 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 knowledge, multiple procedures for measuring that knowledge, multiple criteria for determining 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 knowledge** (general language capacity unrelated to a particular language program) assessed via a wide array of assessment procedures. The multiple approach is implemented in various phases of the assessment 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, motivating, categorizing and exercising power.

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 performance 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 testing repertoire, used either to supplement other tests or on their own. Each of these proce-

Multiple Purposes for 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 a language
- Disciplining learners
- Exercising power in the language classroom
- Conducting research on various facets of language study

dures is chosen on the basis of its characterizing features and suitability for the testing situation and purposes (for example, costs and availability of trained raters).

It is important to note that although tests are not the only means for assessment 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

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 examing the production of oral and written language samples, and include formats such as interviews or essay writing. This division between production and comprehension skills, however, is not always applicable for any kind of task, especially the more comprehensive 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:

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

Multiple ways of designing items and tasks

4) Multiple ways of administering. Rather than the traditional single administration of a paper and pencil test, present 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 for various administration forms are:

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

Multiple ways of administering assessment

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 answer may be one-dimensional as in closed item formats (like multiple choice or matching item types), or open to multiple interpretations as in a performance tasks. 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 (more on designing rubrics on p.). Criteria may also appear in the form of rating scales**,** either **holistic rating scales** (rating scales which assess global language ability) or **analytic rating 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.

- 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

Multiple criteria for determining language quality

Total score
Standards, benchmarks, competencies, can-dos, band scales
Diagnostic criteria
Holistic rating scales
Analytic rating scales
Rubrics
Guidelines (e.g. ACTFL, ISLPR)
Native / non-native criteria

6) Multiple criteria for determining the quality of assessment procedures. Assessing the quality 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 went on prior to the assessment activity.

We distinguish among a number of validity types each relating to a different aspect in the assessment: **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 ability, 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 is intends to assess the higher 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 terms of meaning construction. **Face validity** examines whether there is a match between what he test actually looks 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 actually used and whether 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 designed 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 smaller 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 (like 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 conditions, 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 disagreement 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 are 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 assessment 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 immigrant, 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, results can be reported to various stakeholders – including parents, administrators, 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 parties, 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 assessor 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 an assessment conference (with the student and/or other participants)

Multiple ways of reporting results

- As test scores
- Proving 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 semester 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 administration.

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 although 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 definition 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 consider the ethical and moral questions as well as dilemmas involved in designing and administering 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 consequence of these decisions be that are based on the assessment? Will certain segments of the population be affected 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, Lyle F. 1990. *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L.F. & Palmer, A. S. (1996).). Test Usefulness: Qualities of language tests In *Language Testing in Practice* (Chapter 2). Oxford: Oxford University Press.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second-language teaching and testing. Applied Linguistics, 1 (1), 1-47.

Cicourel, A.V. & Katz, A.N. (Eds., 1996). Special issue on ecological validity in pragmatic research', *Pragmatics & Cognition,* 4, *2*.

Fulcher, G. (2000). The 'communicative legacy in language testing. *System* 28, 483-497

Hymes, Dl. (1974). *Foundations in sociolinguistics*. Philadelphia: University of Pennsylvania Press

International Second Language Proficiency Ratings (ISLPR) - Australia

Kunnan, A. J..(2005) Language assessment from a wider context. In Hinkel, E. (ed.) *Handbook of research in second language teaching and learning*. Mahwah: Lawrence Erlbaum Associates Publishers

Lynch, Brian. (1996). *Language Assessment and Program Evaluation*. CUP

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. *The Forum* . http://eca.state.gov/forum/vols/vol38/no1/p18.htm

Popham, W.J. (1997). Consequential validity: Right concern - wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8, 13

Shepard, L. (2000) The role of assessment in a learning culture . *Educational Researcher, Vol. 29, No. 7, pp. 1-14* . PDF on Professor Shepard's home page: http://www.colorado.edu/education/faculty/lorrieshepard/assessment.html

Shohamy, E. (1998). Evaluation of learning outcomes in second language acquisition: A multiplism perspective. In Byrnes, Heidi (Ed.) *Learning Foreign and Second Languages*. NY: The Modern Language Association of America.

Shohamy, E. (2001). *The power of tests*. Pearson Education Ltd.: Harlow, England.

Spolsky, B. (1975). Language testing – the problem of validation. In L. Palmer & B. Spolsky (Eds.). *Papers on Language Testing 1967-1974.* Washington, D.C.: TESOL.

This CALPER Professional Development Document was produced with funds from a grant awarded to CALPER by the U.S. 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.

Revised October 2010:

URLs were updated

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, CALPER.

Center for Advanced Language Proficiency Education and Research
The Pennsylvania State University
5 Sparks Building
University Park, PA 16802
http://calper.la.psu.edu