Luukka, M.-R., S. Salla & H. Dufva (toim.) 1998. Puolin ja toisin. AFinLAn vuosikirja 1998. Suomen soveltavan kielitieteen yhdistyksen julkaisuja no. 56. Jyväskylä. s. 277-287.

LANGUAGE TESTING: RECENT DEVELOPMENTS AND PERSISTENT DILEMMAS

Sauli Takala

Jyväskylän yliopisto

This article discusses recent developments in language testing. It begins with a review of the traditional criteria that are applied to all measurement and outlines recent emphases that derive from the expanding range of stakeholders. Drawing on Alderson' seminal work, the article then presents criteria for evaluating communicative language tests. Developments in the authentic/alternative assessment movement are briefly reviewed and the merits and limitations of traditional and alternative assessment are compared and contrasted. Some persistent problems in language testing are discussed: methods effect, classification errors, rater agreement, problems of the local independence of items, and a cavalier attitude towards the error of measurement. The article concludes with an optimistic tone: new developments in test theory promise better answers to the perennial problems.

Keywords: language testing, test theory, authentic assessment, error of measurement

1 GENERAL INTRODUCTION

Evaluation is usually considered an activity whose purpose is to determine the worth (merits, quality) of objects, performances or activities, programs or systems. Evaluation needs criteria for what counts as quality (characteristics, attributes of merit). In education - including language education - curricula and syllabi normally function as such criteria. Thus, there needs to be a very close link between objectives and evaluation. Tests are an important, though by no means only, source for making evaluations.

1.1 Evaluation: some key questions

Evaluation can - following Brian North (1993) and others - be regarded as the principled observation of performances in a variety of tasks in order to gather information and to report relevant aspects of that information to interested parties. Testing is one of many possible and useful ways of gathering such information.

This information exchange is facilitated if the assessment procedure is

- open and comprehensible (transparent)
- internally consistent
- can be related to other assessment systems

This view means that all assessment needs to be informed by and based on answers to the following **questions**:

1) What information (Why test? When? What for?)
2) How to organize and report information (What for? Who for?)
3) What to test? (What model of L2 competence? Content? Sampling of content/students?)
4) How to evaluate performance? (Count or judge?)

A number of **audiences** have an interest in what the evaluation outcomes are. Such 'stakeholders' are eg.

- individuals (pupils, students; teachers)
- institutions that provide educational services, programs (schools, universities ...)
- local/district/regional educational authorities
- national educational authorities (Ministry of Education, Parliament)
- international/transnational institutions (OECD/educational indicators; UN; European Union)
- interest groups/lobbies (industry, business; the general public, minority groups; media ...)

The increased interest in what evaluation has discovered is more and more manifested in a demand for **accountability**: decision-makers at various levels are asking for evidence on how effective is teaching in individual schools, at the regional and national level and in the international perspective. The effectiveness and productivity of schooling, the educational field are of concern all over the world.

As the stakes are often high for individuals and for institutions, there are strict **requirements for all assessment/testing:**

- Reliability (intrapersonal and interpersonal agreement on scores, ratings, interpretations)
- Validity (adequate basis/evidence for conclusions, interpretations, judgements; construct, content, concurrent, predictive, ecological; consequential validity)
- Practicality/economy

Validity is the most essential requirement. However, validity presupposes reliability. Strong insistence on perfect reliability/objectivity may, however, lead to validity problems (easy to score discrete points tested by multiple choice, filling of very small gaps, error counting, etc). Thus, if we can enhance reliability in the assessment of more "open" tasks, we also contribute to more valid assessment.

There appear to have been different traditions in language testing (ie., American vs. British tradition): **statistical analysis** of student performances drawing heavily on test theory and statistical analysis vs. **experts' judgements** relying relatively more also on the theoretical and practical experience of constructing tests and marking them.

However, as Alderson (1993) has shown: use of 'experts' is subject to problems:

1) Experts do not agree very strongly on what is being tested (by a question, item..).
2) Experts do not agree very strongly on the difficulty of tasks/questions/items.
3) Experts' revisions - even when based on empirical item analysis data - may not lead to a better test.

Long-term personal experience in language testing suggests that statistical analysis is extremely valuable in judging items, tests, ratings etc. but cannot give simple and straightforward answers. It is a good tool for interpretation, but it cannot be a substitute for subject-matter expertise. Scores, norms, statistical indices etc. need to be critically checked and interpreted by the 'user'.

Testing and evaluation serve so many different needs and audiences that several **types of testing/assessment** have developed over the years, eg.,

- Norm-referenced testing vs. criterion-referenced testing
- Achievement testing vs. proficiency testing
- Diagnostic testing vs. formative testing vs. summative testing
- 'Standardized' tests vs. teacher-made tests
- External vs. internal testing/assessment
- Self-assessment, peer-assessment, teacher-assessment, external assessment
- High-stakes vs. low-stakes assessment
- Tests, examinations vs. national assessments (representative samples)

2 EVALUATING LANGUAGE TESTS

As was noted in the above, language testing is a widespread form of activity, which has many uses and can have important consequences for individuals and groups. Language testing is a fruitful domain for applying new developments in linguistics and applied linguistics, second language acquisition research, psychology and psycholinguistics, sociology and sociolinguistics, discourse/conversation analysis and text linguistics, education, language pedagogy, test theory and psychometrics, and others. Language testing also needs to respond to emerging needs of individuals and societies.

It is evident that in all testing, language testing included, there has been a growing concern with validity issues. Traditionally, validity has been viewed as a question of content appropriateness. One aspect of this concern has been the major attempt to make sure that the test corresponds to what has been taught or what kinds of communication skills are needed in the workplace. Content continues to be one important feature to consider in making and judging validity claims.

Almost twenty years ago Alderson (1981) asked. "How are we to evaluate communicative language tests? What criteria are we to use to help us construct them, or to help us determine their validity?" Alderson asked:

1. What is the test's view of language?
2. What is the test's view of the learner?
3. What is the test's view of language learning?
4. What is the role of background knowledge?

Since Alderson first wrote the above questions, theoretical and empirical research has provided some evidence that helps us to address some of his questions in a more principled manner than before. However, the language testing research and development community needs to work hard for a long time to be able to give good answers to a large set of more specific questions he asked.

3 RECENT TRENDS: TOWARDS 'ALTERNATIVE',

'AUTHENTIC' AND 'PERFORMANCE' ASSESSMENT

Experts on "authentic assessment" tend to agree on a number of points concerning authentic assessment:

- the aim is to assess skills and abilities in contexts that closely resemble the actual situations in which they are used
- assessment tasks are an integral part of studying and learning
- assessment tasks focus attention both on the learning process and its outcomes
- assessment tasks stress the application of knowledge, critical thinking and problem-solving
- assessment tasks put more emphasis on the students' own production than on them answering preset questions (on-demand responding)
- assessment tasks tend to contain large cross-curricular integrated projects rather than separate items
- assessment tasks address not only knowledge but also learning strategies and their monitoring as well as the development of study attitudes
- assessment tasks seek to find out the quality and strengths of learning rather than its quantity and weaknessess

According to Wiggins (1990), one of the chief advocates of "authentic" assessment, "assessment is *authentic* when we directly examine student performance on worthy intellectual tasks. Traditional assessment, by contrast, relies on indirect or proxy 'items' - efficient, simplistic substitutes from which we think valid inferences can be made about the student's performance at those valued challenges.

Wiggins compares traditional standardized tests and "authentic assessment" in the following manner in an attempt to clarify what "authenticity" means when considering assessment design and use:

- Authentic assessments require students to be effective performers with acquired knowledge. Traditional tests tend to reveal only whether the student can recognize, recall or "plug in" what was learned out of context. This may be as problematic as inferring driving or teaching ability from written tests alone.
- Authentic assessments present the student with the full array of tasks that mirror the priorities and challenges found in the best instructional activities: conducting research; writing, revising and discussing papers; providing an engaging oral analysis of a recent political event; collaborating with others on a debate, etc. Conventional tests are usually limited to paper-and-pencil, one-answer questions.
- Authentic assessments attend to whether the student can craft polished, thorough and justifiable answers, performances or products. Conventional tests typically only ask the student to select or write correct responses-irrespective of reasons. (There is rarely an adequate opportunity to plan, revise and substantiate responses on typical tests, even when there are open-ended questions).
- Authentic assessment achieves validity and reliability by emphasizing and standardizing the appropriate criteria for scoring such (varied) products; traditional testing standardizes objective "items" and, hence, the (one) right answer for each.
- "Test validity" should depend in part upon whether the test emulates real-world "tests" of ability. Validity on most multiple-choice tests is determined merely by matching items to the curriculum content (or through sophisticated correlations with other test results).
- Authentic tasks involve "ill-structured" challenges and roles that help students rehearse for the complex ambiguities of the "game" of adult and professional life. Traditional tests are more like drills, assessing static and too-often arbitrarily discrete or simplistic elements of those activities.

It is maintained that a move toward more authentic tasks and outcomes thus improves teaching and learning: students have greater clarity about their obligations (and are asked to master more engaging tasks), and teachers can come to believe that assessment results are both meaningful and useful for improving instruction. If our aim is merely to monitor performance then conventional testing is probably adequate. If our aim is to improve performance across the board then, Wiggins insists, the tests must be composed of exemplary tasks, criteria and standards.

Performance assessment, where test takers have to demonstrate practical command of skills acquired/needed, is more and more commonly introduced to replace or at least complement more traditional test formats, for instance, multiple choice questions or short answers. The relevance of performance assessment is immediately obvious in the context of the workplace. Knowledge of foreign languages is increasingly an integral part of occupational/professional qualifications, and it is expected that occupationally/professionally oriented language tests measure concrete, practical and relevant skills.

It seems obvious that *portfolios* are a promising tool of alternative assessment to be added to the language teachers' methodological toolbox. Properly used they are likely to be beneficial both in learning and the assessment of learning.

Yet, even if "alternative" forms of assessment have certain attractive features, it as well as "traditional" assessment both have some limitations in addition to certain advantages. The following table (drawing especially on Messick 1992) compares alternative assessment with more traditional assessment trying to present a balanced view.

Thus, it appears that there is a trade-off working here. Advantages often are bought at the expense of disadvantages. At all events, if we are aware of the pros and cons, we will be in a better position to make informed choices.

Feature	Aim/goal/intention	Potential strengths	Potential criticisms
Authentic (alternative, performance)	Assessment must reflect a "modern" view of learning and the natural uses and contexts of	* Important and valuable goals are assessed * Assessment is in line with the curriculum and	* Authenticity is not an unequivocal concept and thus does not have unequivocal criteria either

assessment Traditional (multiple choice-based) assessment	knowledge Assessment should, above all, be reliable and commensurate - the context of use is secondary	even supports its attainment * Assessment is felt to be meaningful and motivating * Assessment reflects a person's strengths and may bolster self-image . * Subjectivity is under control * Reliability is generally good * The domain to be assessed is covered well * Assessment is cost-effective	* Alleged benefits of authentic assessment lack strong, solid evidential basis * Validity can be a problem * Washback effect on teaching may be undesirable * Assessment may focus too much on memorization, and larger knowledge structures may be neglected
Degree of directness of assessment (testing): More direct More indirect	Assessment must reflect its target as closely as possible; the effect of target-irrelevant factors should be a minimized	* Face validity of assessment is good * Interpretation of results is more clearcut (low-inference) * Probably a better control of assessment target * More objective scoring	All assessment is indirect and always requires interpretation * Scoring requires 'subjective' judgement (methods variance) * Face validity weaker * Interpretation of results less clearcut (high-inference)
Assessment based on tasks (task-driven)	Enhancing the 'pragmatic' aspect of validity	* Assessment is credible since authentic tasks allow, and require, the use of all important skills and knowledge necessary for a good performance	* It is not easy to define tasks in an unambiguous manner. * It is not clear how generalizable information is obtained by task-based assessment

Assessment based on the cognitive basis of knowledge and skills (construct-driven)	Enhancing the 'conceptual' aspect of validity	* Assessment is generalizable, since it is known what the tasks are based on	* Interpretation is not as straightforward as in task-based assessment
Assessment based on a very open situation Assessment based on a highly structured situation	Enhancing "real-life" linkage Enhancing reliability and control of error	* Assessment corresponds well to "real life" where the situations are often "open" and a person has to decide for him/herself what it is all about * Assessment situation is well under control:diagnostic information is obtained at desired level of accuracy ("grain") * Restricted assessment situation creates a sense of security	* Openness may baffle and lower performance for some individuals * Openness is relative - even partly structured situations may be close to "real life" * Assessment is artificial and does not provide an adequate picture of proficiency * Structured situation may be felt to be too restrictive, which may lower motivation

4 PUZZLES AND DILEMMAS

In spite of - or somewhat paradoxically, because of - more research on testing and assessment and the enhanced knowledge base, there are a number of dilemmas that deserve attention. Below I will list some.

- Test takers may understand something but do not know how to show that they cannot, so to speak, perform their competence. This may distort the test outcome. One example of this is the so-called test-method effect (or bias): the method used in testing may favour some people and disadvantage others. One way of avoiding the test-method bias is to use more than one method of testing a particular skill.
- The fact that a person responds correctly to some item or task does not necessarily mean that he or she actually knows what the item or task is supposed to measure. People may through sheer luck arrive at a correct answer even though they have applied a wrong procedure. This threat to validity can be diminished by measuring the same topic by more than one item/task.
- When a teacher improves teaching, some will benefit but others may be baffled by the new approach and their learning may suffer, at least in the short term. There is increasing evidence in research literature that thinking styles differ. This means that different students should, to the extent possible, be given the opportunity to study - and be tested - in the manner that suits best their thinking style.
- If we double our information about testees (for instance, by using twice as many questions), the error of measurement due to the testees decreases (eg. in the California 1993 assessment, writing, the error variance deminished by 30 % when a second writing task was introduced). However, this may have a paradoxical effect on the evaluation of the quality of an educational programme/system. Good students are not necessarily good on all possible domains of knowledge and skills, and thus a better coverage of the content domain may lead us - questionably - to claim that the standards of the best students have fallen (Cronbach 1995, Appendix: 55)
- It is often suggested that rating (marking, classification) is more reliable if you have only a small number of categories or levels. Classification errors are unavoidable, but if a rater is forced to use only numerical categories (say 1-2-3-4-5), and is not allowed to use a finer classification (say, 1+, 1++, 2-) classification accuracy is likely to suffer: classification accuracry is more reliable in the mid-regions (1,5, 2,5 etc) than close to the category boundaries (Cronbach et al., 1995: 7, 26).
- Psychometric theory presupposes local independence, the independence of elements (items). On the strict interpretation, one can only ask one question

about a reading or listening passage. This might mean that testing of the main idea comprehension is actually the only statistically fully defensible form of testing comprehension. If we, however, believe that comprehension is more complex than that, we may be well advised to treat e.g. text comprehension tests as units (with a mean level of difficulty) rather than as consisting of several independent items. Note that, by contrast, speaking and writing products can be assessed separately on different criteria, without jeopardising the requirement of local independence (Cronbach et al., 1995: 24).
- In assessment, it is often necessary to use raters who use a rating scale (say with levels from 1 to 5 or 1 to 9). Let us assume that the level of perfect agreement between two raters is 60 %, which means that there would be relatively speaking fewer cases where raters differ by only one scale point and even considerably fewer cases with a divergence of two scale points. Sixty per cent perfect agreement sounds quite good, a respectable level of agreement. Let us assume further that the test takers represent a normal sample. Most of the cases would cluster around level 3 or 5, respectively. This means that if one of the raters does not even read or listen to the products but always assigns the middle level score, quite a high level of agreement would appear as an empirical outcome. If, for example, 45 % of test takers receive a grade of 3, a 45 % perfect agreement would be obtained in this manner. A level of 60 % perfect agreement does not sound so very satisfactory if 45 % perfect agreement can be obtained actually by chance (Cronbach et al., 1995: 11)
- Traditional test theory was developed to analyse tests in terms of how much error, or conversely, how accurately differences between individuals can be measured. This is reported by the traditional reliability coefficient. However, the situation is more complex. There are problems if we wish to measure ability in absolute terms, estimating performance against certain criteria and stating what percentage of persons perform at certain levels of proficiency. This kind of measurement, which appears to be spreading, requires the development of appropriate test theory. If we wish to report results at the school level, there are also great conceptual difficulties in terms of reliability estimation, since we are no longer operating at the individual level (on which most theory is based). Cronbach et al. (1995) suggest that computing a standard error is a proper solution to the problem and reporting the confidence band within which the true score can be expected to be found with, say, 95 % level of confidence.
- If our tests or examinations are high-stakes for individuals or schools (ranking of schools/league tables, rewards or punishments) how we deal with the potential attempts by them to beat the system. The schools may even encourage weak students to stay at home on testing days in order to raise the school scores (Cronbach 1995: 7-8).
- There are two main lines of estimating how reliable scores are: Generalizability Theory and Item Response Theory. Generalizability theory considers an observed (empirical) score as a sum of several components: tasks, purposes, classes, schools, students and raters. Item Response Theory (Rasch-model) considers the score as the result of two components: the difficulty of the item and a person's ability. As far as I can judge, the two methods appear to complement each other, to some extent (while they also do partly the same job): generalizability theory seems useful when programmes are evaluated and IRT when individuals are tested.

Undoubtedly other puzzles and dilemmas could be added if we also turned to eg. the social aspects of testing and assessment.

5 CONCLUSION

As an expert in psychometrics, and as one of the main contributors to the development of the powerful new tools in test theory, Cronbach (1995) in his valedictory speech expressed his worry about the neglect of proper attention to the error in measurement. Errors due to sampling of students/classes/schools, errors due to the assessment methods, errors due to rating etc. are not taken into account adequately when assessment /testing is being planned, carried out and reported. This problem is aggravated when new - highly desirable - methods of assessment are being introduced. The situation is not limited to testing/assessment only: whenever anything new is being introduced, we always are relatively speaking novices, trying to learn how the thing can/should be properly done. Novices make a lot of errors.

This means that all assessment/testing should make maximum use of the new solutions that help us in getting a better idea of the potential sources of error. Error as such is not a problem. The real problem is if we are not fully aware of the sources of error, because then we cannot anticipate sources of error and estimate their size.

Error is unavoidable but a responsible tester/evaluator cannot avoid answering the question: Can we live with the error of this magnitude in our scores, our ratings, our interpretations? I believe that testers and evaluators always need to be asking this question for a number of reasons. One pragmatic reason is that others are bound to start asking such questions increasingly in the future.

Thus, the task of the tester/evaluator is not an easy one, or adapting a phrase from Gilbert & Sullivan, the testers' "lot is not an 'appy one". However, making use of new insights and methodologies will make the job more professional. It might be easier not to have all these complications because one could happily live in a fool's paradise, but as Bertrand Russell once said, only a fool would regard it as a paradise.

References

Alderson, J. C. 1981. Report of the discussion on communicative language testing. In J. C. Alderson & A. Hughes *Issues in Language Testing.* ELT Documents 111. London. The British Council, 55-65.

Alderson, J. C. 1993. Judgments in language testing. In C. Chappelle & D. Douglas (eds.) *A new decade of language testing research.* Washington, D.C., TESOL Publications.

Cronbach, L.J. 1995. A valedictory: *Reflections on 60 years in educational testing.* (htpp://www2.nap.edu/htbin/ (also National Academy Press 1995)

Cronbach, L.J., R. L. Linn, R. L. Brennan, & E. Haertel 1995. *Generalizability analysis for educational assessment.* Evaluation Comment. Summer 1995.

Messick, S. 1992. *The interplay of evidence and consequences in the validation of perfor- mance assessments.* ETS Research Report RR-92-39.

North, B. 1993. Transparency, coherence, and washback in language assessment. In K. Sajavaara, R. C. Lambert, S. Takala & C. A. Morfit (eds.) *National Foreign Language Planning: Practices and Prospects.* University of Jyväskylä, Institute for Educational Research, 157-193.

Wiggins, G. 1990. The case for authentic assessment. *ERIC Digest* ED328611.