Pietilä, P. & O-P. Salo (eds.) 1999. Multiple Languages – Multiple Perspectives. AFinLA Yearbook 1999. Publications de l'Association Finlandaise de Linguistique Appliquée 57. pp. 83–104.

RECENT ADVANCES IN LANGUAGE TESTING

Sauli Takala, University of Jyväskylä
Sari Luoma, University of Jyväskylä
Felianka Kaftandjieva, University of Jyväskylä
Maija Saleva, University of Turku

1 INTRODUCTION

This article is based on the presentations held in 1998 at the AFinLA annual conference in Turku. The conference was organized into thematic parallel sessions, and one of them dealt with language testing. The panel of presenters agreed that the overall topic of the session would address recent advances in language testing. Each panellist would cover a topic that was of particular interest to him/her and which would also complement the other presentations. In addition to the contributions presented below, Professor Viljo Kohonen talked about alternative assessment, but he was not able to elaborate it for the present purposes.

Sauli Takala's presentation (section 1) deals with some general issues related to testing and evaluation. Sari Luoma addresses (section 2) the thorny and crucially important question of validity and process of validation. Felianka Kaftandjieva (section 3) discusses item response modelling and presents an overview of its advantages. In conclusion, Maija Saleva (section 4) presents some views related to the strongly emerging movement of oral testing, drawing on her recent doctoral thesis on the topic.

2 SOME GENERAL ISSUES IN TESTING AND EVALUATION

2.1 Testing an instrument of power

Tests are used for a variety of purposes (eg. Shohamy 1992), for instance, to measure students'/learners' knowledge in relation to future tasks that they are expected to perform; to place students/learners in appropriate tracks, streams or classes; to award certificates; to determine whether students are to be promoted or need to repeat a grade; to select those who are predicted to be most suitable for higher education studies; to motivate learners to study hard, learn and perform well; to select people for jobs.

The tests have a strong impact on the lives of individuals, and they have become powerful tools, capable of changing the behaviour of those who are affected by their results – students, teachers, administrators, educational institutions, policy makers etc. Central educational agencies may use tests as potentially effective tools to introduce new curricula, new approaches, and new teaching methods (planned washback effect, test-driven development work).

For testing to be a legitimate activity, it needs to match the exercise of power with a highly developed sense of responsibility for the quality of work. Testing must combine power with responsibility. Error is unavoidable but a responsible tester/evaluator should not avoid facing the question: Can I live with the error of this magnitude in the scores, ratings, decisions or interpretations?

More responsibility can be shown, in concrete terms, by **always** reporting the standard error of measurement, reporting the confidence band within which the true score can be expected to be found with, say, 95 % level of confidence. This is often easy enough, but there may be some problems if we wish to measure ability in absolute terms, estimating perfor-mance against certain criteria and stating what percentage of persons perform at certain levels of proficiency. This kind of measurement, which is spreading, requires the deve-lopment of appropriate test theory.

2.2 From concept to operationalization

The idea of test specification is not new – some explicit models were applied already in the early part of the century. It is more recently, with the

criterion-referenced movement and the increased sophistication in the definition of validity in test theory, that construct definition has been receiving increased attention. Samuel Messick (1989, 1994), in particular, has drawn our attention to problems of construct-irrelevant variance and construct under-representation in our measures (scores). Constructs are, however, far from easy to pin down.

Let us take the level of proficiency/skill as an example. Drawing on Becker (1998), it might be suggested that there are not groups of people at different levels of proficiency and that there are not levels of proficiency only because of measurable or observable differences from other groups and levels. A possible complementary aspect of the situation is the fact that people talk and act as if there were such groups and levels and they classify themselves as beginning, intermediate or advanced language learners/ users. In other words, perceptions of proficiency may have some relevance, in terms of construct definition, in addition to measured proficiency. Yet, we must be absolutely clear about the fact that perceptions of proficiency/skill are not the same thing as measured proficiency/skill.

Obviously, language proficiency, language skills etc are **concepts**, which as such are very general and vague. For teaching and testing purposes they need to be refined and to acquire more detailed and explicitly defined meaning, in other words they must be developed into **constructs**. Language proficiency, language skills etc, as properly defined constructs, are sufficiently elaborated that instruments (tests, measurement procedures) of different kinds can be developed. These are **operational definitions** of the concept and construct of language proficiency. An orderly/-logical progression and continuity from concept through construct to operationalization is required to ensure construct validity (see, e.g. Black 1999).

We might ask: what kind of constructs do we have and where do they come from? It seems to me obvious that we do not possess very well developed constructs in language teaching and testing. However, there is a more than 2000-year tradition of language teaching to draw upon, and it goes without saying that generations of language teachers/tutors have always done some things in a sensible manner even if there have always been complaints and pressures to reform. Similarly, even if there are no well-developed constructs, there are e.g. useful skill taxonomies, which draw on past wisdom and increa-singly on research findings.

2.3 Multiple forms and audiences of testing and evaluation

Like love, testing/evaluation is "a many splendour thing." It may be of many kinds and it is of interest to many parties. There may be varying, or even conflicting views of what are the main goals and uses of testing/evaluation. We need to realize that there is a trade-off operating here: we cannot serve all purposes well without an extremely complex and expensive design. We always need to prioritize goals/uses and be aware of what the advantages and disadvantages are in each case.

Several 'stakeholders' are potentially interested in the outcomes of tes-ting/evaluations, including e.g.

- 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; 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 teaching is in individual schools, at the regional and national level and in the international perspective. The effectiveness and productivity of schooling, the educational yield, are of concern all over the world.

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

- 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
- High-stakes vs. low-stakes assessment
- Tests, examinations vs. national assessments (representative samples)
- Self-assessment, peer-assessment vs. teacher-assessment

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.

3 RECENT DISCUSSIONS ON VALIDITY AND THE PROCESS OF VALIDATION

3.1 Introduction

Validity is one of the most central concepts in the philosophy of educational measurement. It is associated with such value-laden fundamental concepts as meaning, truth, and worth, which mankind has always found interesting. Validity studies in measurement contexts focus on the meaning of the measure.

As a technical term, validity is one of the two big measurement qualities of reliability and validity. Validity is a quality in two senses of the word. It is a quality as in a property of something, like the height of a building or, perhaps more to the point, the suitability of a building for a particular use, such as providing a home for a family, or holding a conference. The second sense in which validity is a quality is that it is a standard for high quality. Validity is something valuable. It has value in a continuous way: it is about more or less, not about all or nothing.

The concept of validity in educational measurement has evolved dramatically during the past 50 years. It used to be one of many qualities that tests have, now it is viewed as the main one. In the broadest senses, validity is considered to be a higher-order term, such that validation offers a framework for all research that concerns a test. This is because ultimately, each and every technical investigation of any aspect of a test or its scores is connected with the meaning of the measure.

It used to be thought that what needed to be valid was the test. This was particularly the case before the 1950s, which is now known as the pre-theoretical era of validation research. The task of a test developer was to develop a good test, show that it is valid, and that is all, quality is guaranteed. This kind of validity was considered timeless, so that validation only needed to be done once per test, and after that it was appropriate to describe the test as "valid". Validity did not encompass the way that the

scores were going to be used, because test use was a different matter from the quality of a test. Today, in contrast, validity is seen to be ultimately connected with score use. Validity is a property of the **inferences** that people draw, and **actions and decisions** that people take, on the basis of test scores. Validity is grounded; it depends in part on the situation in which the test is developed and used, and the use to which the scores are put. If there is old evidence for the validity of inferences and decisions for **some** purpose in **some** situations, the relevance of this evidence for each new use of test scores has to be demonstrated. Because of this tie with score use, it is not just the test developer who is responsible for validation work; test users too must bear their share of the responsibility for how they use tests.

Current validation practice stresses the importance of **evidence**, so much so that validation cannot be done without it. Evidence can come in the form of theoretical rationales for certain meanings and interpretations, and in the form of different kinds of empirical evidence. The empirical evidence can consist of either words or numbers. The more lines of evidence that support an interpretation of the scores, the better. If the lines of evidence conflict, i.e. if some evidence supports a desired interpretation and some does not seem to be relevant or seems to conflict, this is considered highly useful as well. Such situations lead test developers and users to reconsider the aim of the test and the meaning of the scores, and that is exactly what should be happening. Score meaning is what modern validation inquiries are all about.

The new, broad concept of validity and validation means that the scope for validation studies has become broader than it was in the 1950s. The test is still one of the entities which has to be investigated in a validation program because what the test contains does influence the meaning of the scores. But so do the processes that the test takers go through in performing the test, the content of the assessment scales and the procedures of the scoring, the information given to users on the score categories, and the use to which the scores are put. Figure 1 illustrates the contrast between the pre-theoretical and the modern views of the objects of validation.

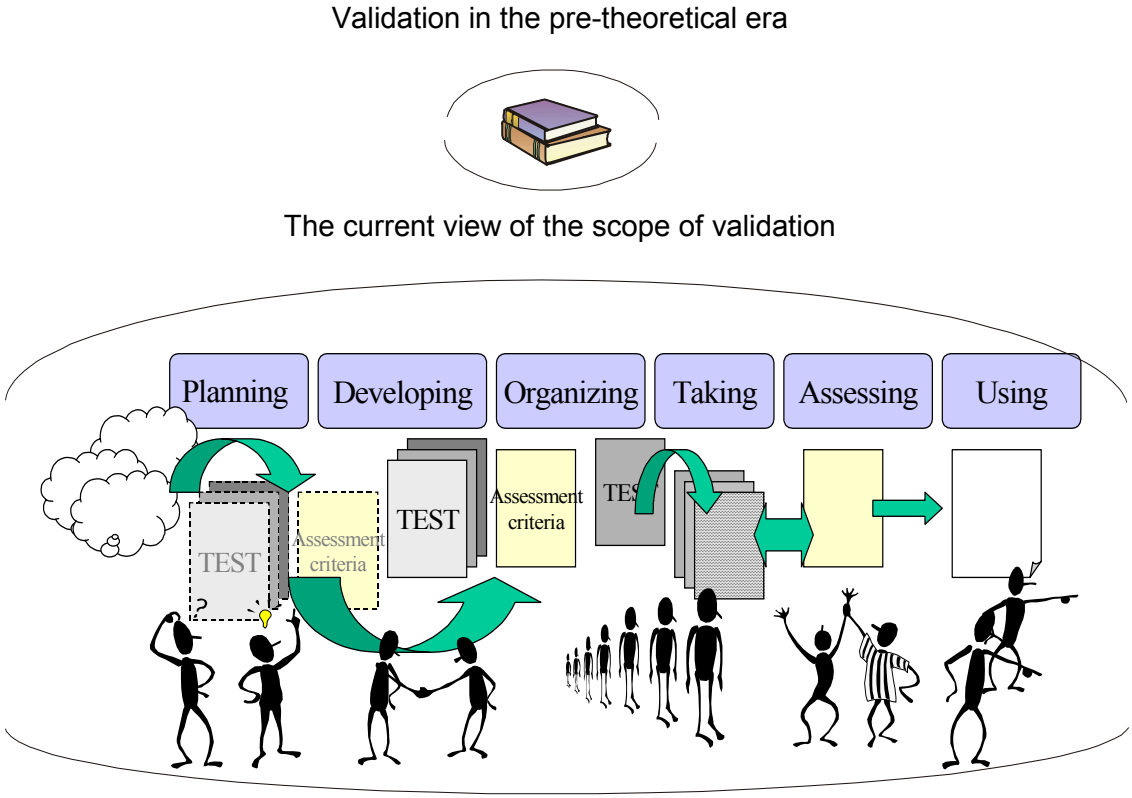


FIGURE 1. Views of the scope of validation

3.2 An activity-based view of testing

Figure 1 can also be used to point out an important difference between the 1950s and the current views of the nature of language tests. The old view was that a test was an object on paper. It consisted of well-structured tasks, which were given in order to get a sample performance of the test takers. Tests and scoring rubrics were developed as they are now, and the content of the test was carefully planned. Little research attention was paid to anything else, however, than the product of the test development process, i.e. the test. This was also what was validated.

To parallel the modern, activity-based and communicative views of language use, language tests have begun to give more emphasis to performance. The tasks in tests require the test takers to produce limited and extended stretches of language, and the processes which the test takers go through when they perform the test tasks are considered to be important. Test taking is seen as a significant activity.

Similarly, test development and administration are also seen as activities. The string of activities which is needed before scores from language tests can be used for some purpose is in fact quite long: the test

needs to be plan-ned, developed, and organized so that it can be taken by the test takers. Their performances must then be assessed, and the results categorized and reported so that they can be used.

All of these activities involve **actors** (developers, organizers, test takers, assessors, users) who do something (create, organize, perform, assess, interpret). Information from all these activities and all these people feeds into the explanation of what the scores mean, and how they can and cannot (could or should not) be used. The activities are partly controlled by test developers, partly by test users. All the activities must be monitored for test development, and all of them also provide data for validation. With this kind of an activity-based approach to language and testing, validation also becomes an activity and a process. The process begins when the idea for a particular test is born and ends when the scores from the test are no longer used for any purpose at all. This is what it means when educationalists say that validation is a framework for all investigations which have to do with score meaning.

3.3 Defining validity

Technically, the broad definition of validity has been put in the following, much quoted format:

> Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. (Messick 1989: 13)

The quote is by a top specialist in educational measurement, but it is very often used by language testers as well. The broad thinking expressed in the definition also guides many validation studies that are conducted in language testing today. Thus, validation studies encompass not just score meaning but also intended score use, and possible unintended side effects.

3.4 The centrality of construct validation

Many researchers, educational measurement people and language testers alike, use the terms validation and construct validation interchangeably.

Over the years, construct validation has indeed become the core concept in validation, as it is this term which always referred to the meaning of the measure. When Wainer and Braun (1988: 1) introduce William Angoff's article, which describes the development of validation as a concept in educational measurement, they point out that "[I]t is particularly interesting to watch the evolution of construct validity, from of a junior member of a troika including criterion validity and content validity, to a coequal status and, finally, to representing 'the whole of validity from a Scientific point of view'. Other types of validation become aspects or strategies in the validation argument". The quote within the quote is from the 1985 version of the Joint technical standard of American Psychological Association (APA), American Educational Research Association (AERA), and National Council for Measurement in Education (NCME), which in the American context is the most influential source of quality criteria for educational and psychological tests. The Joint technical standards, called the **Standards** for short, are used for example in litigation when cases are brought against testing boards or test users for suspected misuse of tests. They are also the basic reference for test development boards when they set quality standards for their development work.

In the introduction to a collection of papers from the 1992 Language Testing Research Colloquium, a volume with the title "Validation in language testing", Alister Cumming (1995: 5) shows that the broad view of validation quoted above has taken hold in language testing: "Rather than enumerating various types of validity ... the concept of **construct validity** has been widely agreed upon as **the** single, fundamental principle that subsumes various other aspects of validation ... relegating their status to research strategies or categories of empirical evidence by which construct validity might be assessed or asserted". Similarly, when Bachman and Palmer (1996: 21) list the essential properties of tests, they mention and define **construct** validity: "Construct validity pertains to the meaningfulness and appropriateness of the **interpretations** that we make on the basis of test scores."

Anastasi (1986: 4–5) defines the kinds of constructs which test development and validation are concerned with as "theoretical concepts of varying degrees of abstraction and generalizability which facilitate the understanding of empirical data". These 'constructs' also often reflect the way that people categorise and think about a test. Thus, when we talk about intelligence tests, listening tests, or driving tests, the constructs here are the mental concepts of intelligence, listening, and driving. People tend to presume that the constructs behind the tests are real, they exist. People

also tend to think that broadly speaking, they agree what intelligence or listening or foreign language proficiency are without having to discuss the details.

For the purposes of test development and validation, the content of the construct is provided by the theory of the construct, which the test developers subscribe to. Depending on how the test was developed, this theory may be more or less precisely formulated, but the theory is nevertheless the basis for the meaning of the construct and hence the meaning of the measure. This meaning is mediated by the way in which the construct was operationalised into the test and its scoring mechanism, and one and the same construct can be operationalised in many different ways. If criticism, **with evidence**, is brought against the way in which someone uses the scores from a language test, this means that something has to change. Quite often, the changes concern the test and its scoring mechanism, or the way that the scores are interpreted and used.

Sometimes, however, changes may also be needed in the underlying theory of language ability.

3.5 Importance of evidence

Evidence is important for validation studies because it helps focus discussions and because it is something concrete. Test developers, validators, and test users can come back to a set of evidence to check interpretations, compare against other evidence, and combine sets of evidence to strengthen old cases or make new hypotheses. Evidence for validation can come from:

- **Documents** related to, and produced in connection with, test development, test administration, test taking, scoring, results analysis, reporting, and use
- **Activities** which constitute test development, test taking/administration, scoring, analysis, reporting of results, result use
- **People** who are involved in these activities

These kinds of evidence offer material for both logical/theoretical analysis and numerical or textual analysis.

The most obvious kinds of data associated with testing are test performances and scores. In addition to these, validation data in language testing contexts can be obtained through judgement, verbal protocols (introspection and retrospection), observations, and interviews (Banerjee & Luoma 1997). All of these have been used in language test validation

studies, and all the stages of test development have been investigated. The assessment stage and the use of scores have mostly been the focus of quantitative studies through analysis of scores. This has recently begun to be balanced with qualitative studies into rating strategies, and interviews with score users about the expectations that they have of test scores.

Not only are there multiple sources of evidence and many different kinds of data which can be gathered from these sources for validation studies, the data can also be approached with several different strategies of analysis. The validators must build a case out of the data that they have, and this begins from categorizing and characterizing the data in some way. The validators' approach to the data can be ethnographically flavored, which entails allowing the data categories to arise from the data itself. Such an approach is common with think-aloud data and intro- and retrospective interviews. A set of data can also be approached with an existing framework which is considered useful for validation purposes. For instance, data on the tasks which a language test contains can be analyzed with the Bachman & Palmer (1996) Task Characteristics framework to compare language use on the test with language use in non-test situations, or to compare the expectations that test developers have for discourse with test discourse as it actually happened with some test takers.

In yet a different approach, the data might be analyzed against predictions from one or more theories. Thus, questionnaire data could be analyzed not only in terms of categories arising from the data, but also in terms of a theory on test anxiety and a theory on how people who want to give a good impression of themselves would respond to the questionnaire. This last example illustrates that triangulation is possible even if an analyst only uses one principal source of data, e.g. test takers, and only one main method of data gathering, e.g. questionnaires. The validation case becomes stronger, however, the more lines of evidence the researchers can show for supporting the validity of the desired score interpretations.

When validation is seen as broadly as it has been outlined above, the task is endless. In that it reflects any human pursuit of excellence. Rather than lamenting the impossibility of completing a validation process, this can be seen as a challenge in somewhat similar ways as a marathon is a challenge for a runner. However, the special nature of testing as an activity makes it a special kind of challenge, one concerned with making a case in the manner of legal cases:

> Because evidence is always incomplete, validation is essentially a matter of making the most reasonable case to guide both the current use of the test and current

studies, and all the stages of test development have been investigated. The assessment stage and the use of scores have mostly been the focus of quantitative studies through analysis of scores. This has recently begun to be balanced with qualitative studies into rating strategies, and interviews with score users about the expectations that they have of test scores.

Not only are there multiple sources of evidence and many different kinds of data which can be gathered from these sources for validation studies, the data can also be approached with several different strategies of analysis. The validators must build a case out of the data that they have, and this begins from categorizing and characterizing the data in some way. The validators' approach to the data can be ethnographically flavored, which entails allowing the data categories to arise from the data itself. Such an approach is common with think-aloud data and intro- and retrospective interviews. A set of data can also be approached with an existing framework which is considered useful for validation purposes. For instance, data on the tasks which a language test contains can be analyzed with the Bachman & Palmer (1996) Task Characteristics framework to compare language use on the test with language use in non-test situations, or to compare the expectations that test developers have for discourse with test discourse as it actually happened with some test takers.

In yet a different approach, the data might be analyzed against predictions from one or more theories. Thus, questionnaire data could be analyzed not only in terms of categories arising from the data, but also in terms of a theory on test anxiety and a theory on how people who want to give a good impression of themselves would respond to the questionnaire. This last example illustrates that triangulation is possible even if an analyst only uses one principal source of data, e.g. test takers, and only one main method of data gathering, e.g. questionnaires. The validation case becomes stronger, however, the more lines of evidence the researchers can show for supporting the validity of the desired score interpretations.

When validation is seen as broadly as it has been outlined above, the task is endless. In that it reflects any human pursuit of excellence. Rather than lamenting the impossibility of completing a validation process, this can be seen as a challenge in somewhat similar ways as a marathon is a challenge for a runner. However, the special nature of testing as an activity makes it a special kind of challenge, one concerned with making a case in the manner of legal cases:

> Because evidence is always incomplete, validation is essentially a matter of making the most reasonable case to guide both the current use of the test and current

> research to advance understanding of what the test scores mean. (Messick 1989: 13)

In a similar spirit, Banerjee & Luoma (1997: 286) conclude an overview of qualitative techniques in test validation by saying: "What the current broad concept of validity and its focus on interpretations and uses highlight is that awareness of what affects meanings and interpretations is a crucial requirement of understandings and meanings. Validation, as it is now viewed, calls for a thorough understanding of the test: the way in which it is constructed; how test-takers give their responses; how these responses are evaluated; and, how the scores are used in making decisions about the test-takers."

4 ITEM RESPONSE MODELLING

In the last decade, Item Response Modelling has emerged as a preferred approach to test development in all fields of educational measurement. The basic reasons for this steadily growing interest in Item Response Modelling are not only its advantages over the Classical Test Theory but also the availability of a variety of user-friendly software products designed for this kind of analysis.

The main advantages of Item Response Modelling have been well known in the psychometrics community for more than 30 years and they can be summarized as follows:

4.1. Sample-free estimation of item parameters

This means that the estimation of such item parameters like difficulty, discrimination and guessing does not depend on the sample data are coming from. Figure 2 illustrates[3] this basic advantage of Item Response Modelling over Classical Test Theory.

One and the same set of items was analyzed using both Classical Test Theory and Item Response Modelling. The left-hand chart presents a

---

[1] The data for this analysis are taken from two sub-samples with almost the same sample size (n1=161 and n2=162), taking part in Yleiset Kielitutkinnot (English–Spring–1996–intermediate level). The scales in this Figure as well as in Figures 3 and 4 represent θ-values.

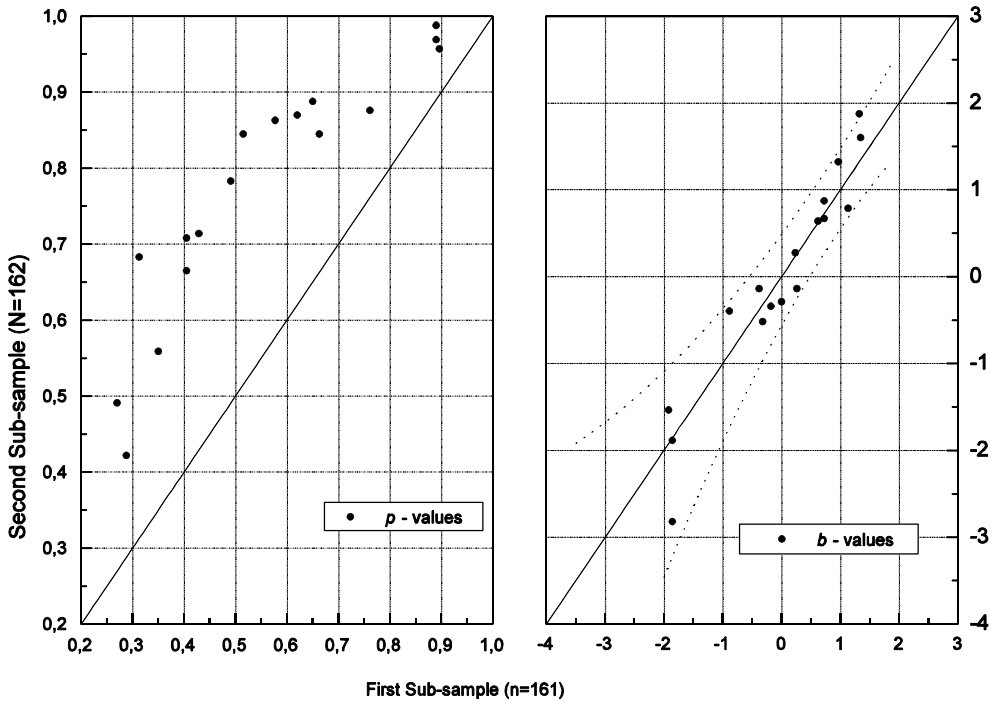


FIGURE 2. Item Parameter Estimation (p & b)

scatter-plot, comparing the item difficulty parameters estimations in terms of p-values (proportion of correct responses). It is quite obvious from the graph that these two estimations, based on different sub-samples, differ significantly. In other words, we cannot make any inferences about the second sample, on the basis of the results in the first sample.

On the other hand, if we estimate the difficulty of the same items in the same sub-samples with the means of Item Response Modelling (b-values) and compare them (Figure 2 – right-hand scatterplot) we can easily see that they are very similar. All they are located very closely to the identity line and are within 95% quality control lines (Wright & Stone 1979: 94–95).

The main application of this feature of Item Response Modelling (the invariance of item parameter estimation) makes it possible to build Item banks and to add new items to the existing Item banks.

4.2 Test-free estimation of person parameters

This means that the estimation of an examinee's language proficiency does not depend on the choice of the items with which the examinee was tested. Figure 3 illustrates[4] this feature. The scatterplot presents the comparison between person estimations, based on two different sub-sets of items (each consisting of 27 items). There is a high correlation between both estimations (0.89), but what is even more important in this case, is that the difference between the two estimations of language proficiency is not significant in 872 out of 900 cases. In other words, 97 % of the person estimations are invariant (stable) and consequently – test-free.

This advantage of Item Response Modelling is broadly used in test equating and computer adaptive testing.

4. 3 Prior information about the standard error of measurement

In Classical Test Theory, the standard error of measurement is a function of test reliability, which is group-dependent (Hambleton & Swaminathan 1985: 123). This means that one and the same test can have different reliability in different samples and consequently the standard error of measurement also is group-dependent.

[2] The data for this analysis are taken from a simulation study (n=900) for Finnish tests in Structure, developed as part of an EU project called DIALANG.

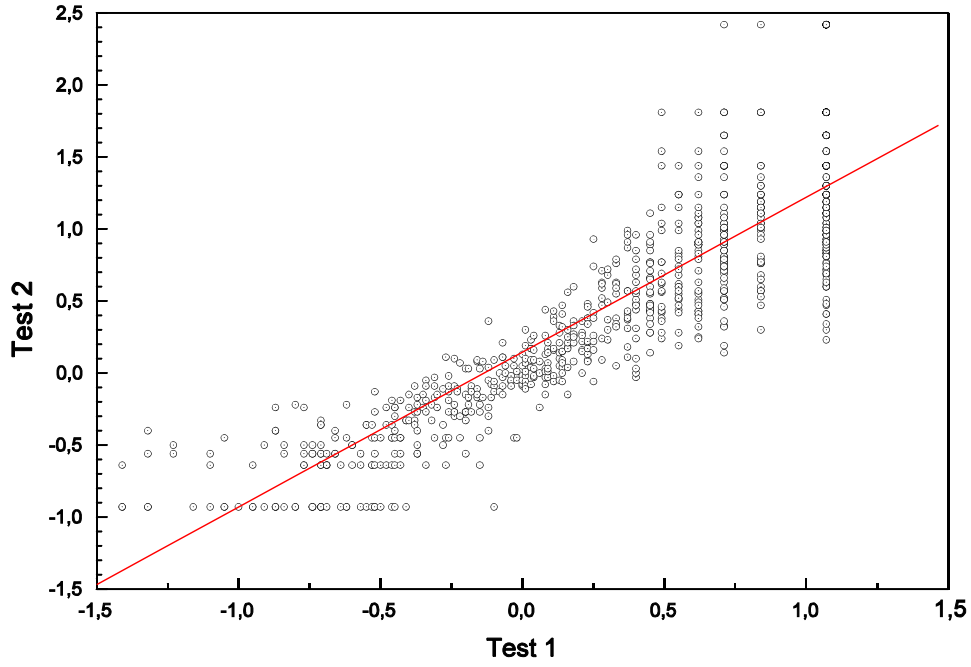


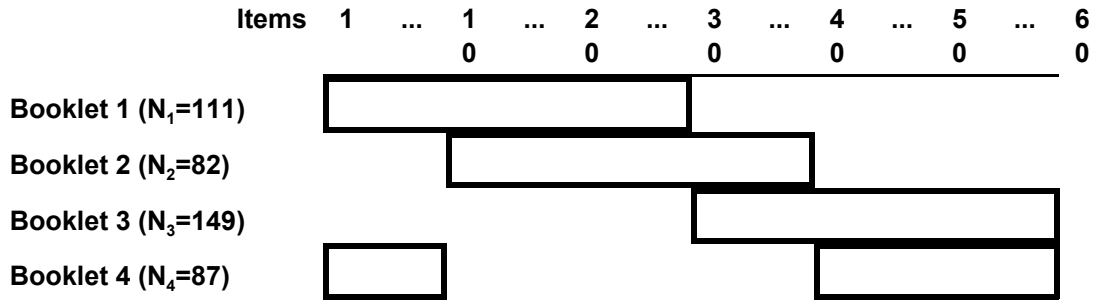
FIGURE 3. Invariance of Person Parameters

On the other hand, the standard error of measurement in its classical interpretation is an average of the errors over different ability levels, but reporting a single number masks this and leads to the wrong conclusion that the error is one and the same on different points of the proficiency scale.

In contrast, the Item Response Modelling makes it possible to determine the standard error of measurement for each point of the measurement scale and it can be used for test design and development, especially in case if the target group characteristics are known beforehand. The following example[5] illustrates the utility of this approach.

In order to develop a computer adaptive test in Finnish Grammar/ Structures, 60 items were piloted with 429 examinees on the basis of linked test design with four booklets, each consisting of 30 items as follows:

[3] The results, presented here, are part of the analysis, conducted for the Prototype development of DIALANG.



Fifty-seven of these sixty items fit the One Parameter Logistic Model (Verhelst & Glass 1995: 215–237). Based on the estimations of item and person parameters – two new tests (each one consisting of 27 items) with a different level of difficulty were designed. The first test (Test 1) includes the easiest 27 items and the second test (Test 2) includes the most difficult 27 items. Although these two sub-tests were never tried out, with the means of Item Response Modelling it is easy to predict their reliability (.95 for Test 1 and .97 for Test 2) and the standard errors of measurement (Figure 4).

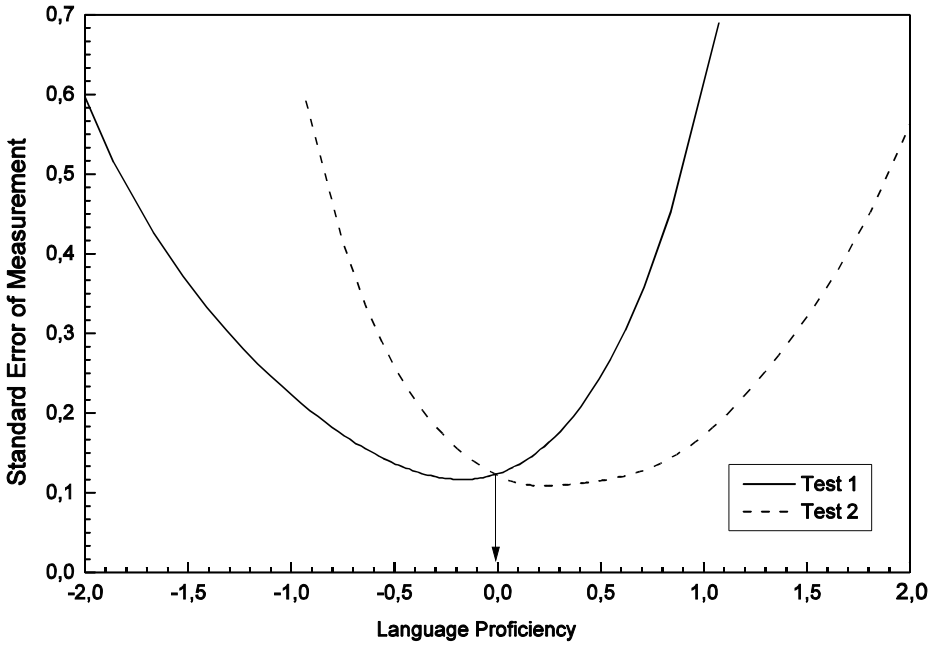


FIGURE 4. Test Precision – Prior Estimation

It can be easily seen in Figure 4 that both tests are almost equally precise, but in different intervals of the scale. While Test 1 is more precise in the interval
(-2; 0), Test 2 is more precise in the interval (0; +2). In other words, Test 1 is more appropriate for clients who are on lower levels A1, A2, B1 of language proficiency, according to Council of Europe proficiency scales, and Test 2 is better for testing higher levels of proficiency (B2, C1, C2).

These are not the only advantages of Item Response Modelling, which make it a highly valuable approach in language testing. However, apart from all its attractive features, it has some strong limitations, which have to be taken into account.

The first limitation is that the stable estimation of item and person parameters can be reached only with comparatively large samples of items and examinees. According to Hambleton (1993: 171) the minimum number of items is 20 and the minimum number of examinees is 200, and this applies to one-parameter model only. In case of two- and three-parameter models, the sample sizes grow up to 60 items and 1000 examinees. Otherwise the main advantages about the invariance of item and person parameters will be violated.

The second limitation is that in comparison with Classical Test Theory, Item Response Theory has very strong and restrictive assumptions, and in many cases, it is very difficult to meet them. That is why a study of the model-data fit is the first and obligatory part of any application of Item Response Modelling and this study should not limit itself only to statistical tests of fit, but should go further to checking model assumptions, expected model features and predictions (Hambleton 1993: 172–182). The reason for this requirement is that in case of misfit (which unfortunately is not uncommon), all inferences and conclusions would be wrong or worthless.

The third purely technical limitations are that Item Response Modelling usually requires powerful computer equipment, appropriate software, advanced statistical knowledge and some preliminary experience. Summarizing, Item Response Modelling is for those who prefer quality and are prepared to pay the price.

5 TESTING ORAL SKILLS – A TIMELY AND IMPORTANT CHALLENGE

5.1 Oral skills need upgrading

One of the problems of language teaching in Finland as well as in many other countries is the fact that too little emphasis is laid on teaching the oral skills, particularly speaking. On and off, this deficiency has been pointed out by various writers and some remedies have been suggested, but the fact remains that when the very newest report on the state of Finnish language education was published (Huhta 1999), the three problem areas were found to be too narrow range of languages studied, insufficient oral skills, and deficient intercultural competence. The latter two can actually be regarded as expressions of the same weakness: in the written form of language use the cultural differences are much fewer and in a writing situation the student has much more time to plan and revise his message paying attention to also cultural conventions. It is the spoken message that is mainly affected by the intercultural incompetence.

It is the present writer's opinion that the main reason for the inadequate emphasis on the speaking skill is the fact that speaking is very rarely tested at any stage of language learning, and most important of all, it is not tested in the national matriculation examination. This state of affairs may be due to factors such as teachers' and students' general resistance to change, fear of a greater workload, teachers' feelings of the insufficiency of their own speaking skill, and old-fashioned language teacher education. The commonest claim of those opposed to an oral test has been the argument that it is a practical impossibility to carry out a large-scale speaking test.

5.2 A specific test is needed to test speaking

To find out whether such an argument was justified, the writer – about ten years ago – started a large-scale research project, which developed into a doctoral dissertation. The main aim of the study was to find an answer to the question whether it was possible to assess senior secondary school students' oral English proficiency in a valid, reliable and efficient way.

When discussing the feasibility of an oral part in the matriculation examination, it has been suggested that only a small percentage of stu-

dents should be tested each year. That is naturally a possibility and maybe even a good alternative to make a beginning, but the writer was guided by the hypothesis that it would be possible to test the whole population of students, about 30,000 students. If it could all be done on one day, as is the case with the present matriculation tests, it would mean a considerably greater chance of reliable measurement. For this purpose the only alternative was a test in the language laboratory. The test that was developed was called LLOPT (Language Laboratory Oral Proficiency Test).

Before proceeding to discuss the developed test, it is worth considering another view which questions the case for large-scale oral testing. It might be suggested (and the topic has been addressed in a doctoral dissertation in Finland by Hellgren 1982) that oral proficiency could be tested with sufficient adequacy indirectly, through the written medium. A review of the domain of speaking showed, however, that there are reasons to believe that spoken language is in important respects different from the written language.

There are two factors that particularly affect the production of speech and make the conditions of speaking different from writing. The first one is related to the internal conditions of speech, the fact that speech takes place under the pressure of time (the processing condition). The second involves the dimension of interpersonal interaction: the fact that two or more people are engaged in social interaction (the reciprocity condition). Furthermore, one of the aspects that distinguishes oral discourse from written is the conspicuous presence of non-verbal communication (including pronunciation and prosodic features). In sum, it seems quite obvious that speaking skill can validly be tested only by means of a test of speaking.

5.3 Interview tests and tests taken in the language laboratory correlate well

The analysis of the domain of speaking led to a conclusion that a valid test needs to be many-sided. It is not enough to assess only one component or to use only one test format or one rating criterion. The LLOPT test included, in fact, five subtests, based on a unified thematic framework: meeting and interacting with American visitors, a youth orchestra, in Finland and repaying the visit later on. The subtests were: reading aloud a letter, interpreting L1 questions in L2 (to help one's mother who wishes to ask a visiting American student about different things), retelling a story in

the local newspaper to the visitor, giving a talk on the Finnish education to the visiting group, reacting in situations and expressing opinions. Each test was rated using criteria adapted to the subtest. The average inter-rater agreement among three independent raters, trained by the author who was one of the three raters, was .96.

The writer obtained responses from a total of 60 students in two different types of senior secondary school. Both schools had taken part in a national experiment with intensified teaching of oral skills. For validation purposes, the writer first administered an ACTFL OPI interview to all students, after having been trained in the use of the system by instructors from ACTFL who taught a course in Finland. The LLOPT test produced an average of 20 minutes of student speech. Data collection succeeded well, and the students were highly motivated.

The results indicated that the LLOPT test correlated highly with OPI (.88), which supports the earlier studies about successfully simulating the interview in a language laboratory (SOPI). The test correlated also strongly with the school report mark (.83) and the matriculation examination grade (.80). In the matriculation examination, the highest correlation at the subtest level was with free essay (.80), indicating that productive skills correlate quite well in the student population. Yet, the results indicate also that only about 65% of the variance in speaking scores can be predicted by the essay task or by the whole test battery in the matriculation examination. One third of the variance is not accounted for. The oral test also placed the students in a different order. Only 46% received the same grade as they did in the matriculation examination. This lends empirical support to the theoretical view that oral skills need to be tested by a speaking test, in order to be sufficiently reliable and valid.

5.4 A valid test is there – why not use it?

However, the question about practicality and efficiency has not yet been addressed. For the purposes of the study, the test was on purpose made versatile and the criteria were also many and varied. A regression analysis showed that the overall score could be predicted effectively with two subtests only: in the present study, the best predictors were responding in situation and interpreting the questions to the foreigner. The first one accounted for about 85% of the variance in the total score and the latter added another ten per cent to the variance explained. This means that

about 10 minutes speech sample is enough for obtaining a valid rating, and the tasks can be few in number.

It can be concluded that the language laboratory can be used effectively to test speaking. This standardizes the situation and enhances reliability, which is a key concern in any large-scale testing. To put this know-how in use requires that schools have well-functioning language laboratories and that teachers are properly trained. Like in all innovations, the testing of oral skills should be introduced without too much ambition – the strategy of starting small is desirable. In the beginning, perhaps speaking might be linked with listening comprehension.

In conclusion, we might carry out a thought experiment: What would present-day in schools and the students' learning profile be like if the matriculation examination only consisted of the translation examination, as it did some thirty years ago? The need of improved oral skills is obvious and testing oral skills would certainly promote the teaching and learning of them. We know how oral skills can be tested in a sufficiently reliable and valid manner, and also economically. The time for concerted action has come, so let us do it.

6 CONCLUSION

In this brief co-authored article, a few topics of current interest in (language) testing were addressed. It was argued that oral testing needs to be made a regular part of all testing in order to support the teaching and learning of a much needed and highly valued skill. It was shown that oral testing can be carried out with sufficient reliability and validity, and it can be done efficiently on a large scale, provided that proper training is arranged for teachers. It was also shown how validity can now be addressed in a sophisticated manner and how validation of tests and testing procedures can be addressed in a principled fashion. Increased awareness of validity concerns and implementation of more systematic validation procedures will improve the quality of language testing. Developments in test theory (item response modelling) also provide new powerful tools that will assist not only the analysis of results but also the planning of new tests. When the new developments are applied, it is possible to predict quite accurately how reliable the new test will be. This is highly desirable because testing can have a powerful influence on people's lives and life opportunities. Testers must be aware of their responsibility for ensuring that testing is carried out with very high quality. They must recognize that there is always error in

measurement, take all possible steps to have a good estimate of the size of error in advance, and then decide if they can live with error of that size. Testing needs to combine professional skill with social responsibility.

References

Anastasi, A. 1986. Evolving concepts of test validation. *Annual Reviews of Psychology* 37, 115.

Bachman, L. & A. Palmer 1996. *Language testing in practice*. Oxford: Oxford University Press.

Banerjee, J. & S. Luoma 1997. Qualitative approaches to test validation. In C. Clapham & D. Corson (eds.) *Language testing and assessment, Vol. 7 of the Encyclopedia of language education.* Dordrecht: Kluwer Academic Publishers, 275–287.

Becker, H. S. 1998. *Tricks of the trade. How to think about your research while you're doing it.* Chicago: Chicago University Press.

Black, T. R. 1999. *Doing quantitative research in the social sciences*. London: Sage.

Cumming, A. 1995. Introduction: The concept of validation in language testing. In A. Cumming & R. Berwick (eds.) *Validation in language testing.* Clevedon: Multilingual Matters, 1–14.

Hambleton, R. K. 1993. Principles and selected applications of Item Response Theory. In R. Linn (ed.) *Educational measurement*. Phoenix: American Council on Education and Oryx Press, 147–200.

Hambleton, R. K. & H. Swaminathan 1985*. Item Response Theory: Principles and applications.* Boston: Kluwer-Nijhoff Publishing.

Hellgren, P. 1982. *Communicative proficiency in a foreign language, and its evaluation.* Research report 2. Department of Teacher Education. Helsinki: University of Helsinki.

Huhta, M. 1999. *Language/Communication skills in industry and business*. Report for Prolang/Finland. Helsinki: National Board of Education.

Messick, S. 1989 Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher* 18 (2), 5–11.

Messick, S. 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23 (2) 13–23.

Messick, S. 1989. Validity. In R. L. Linn (ed.) *Educational measurement.* Third edition. New York: American Council on Education / McMillan, 13–103.

Saleva, M. 1997. *Now they're talking. Testing oral proficiency in a language laboratory.* Studia Philologica Jyväskyläensia 43, Jyväskylä: University of Jyväskylä.

Shohamy, E. 1992. New modes of assessment: The connection between testing and learning. In E. Shohamy & A. R. Walton (eds.) *Language assessment for feedback: Testing and other strategies*. Johns Hopkins University: National Foreign Language Center, 7–28.

Wainer, H. & H. Brown 1988. Historical and epistemological bases of validity. In H. Wainer & H. Brown (eds.) *Test validity*. Hilsdale, N.J.: Lawrence Erlbaum.

Verhelst, N. D. & C. A. W. Glas 1995. The One parameter logistic model. In G. Fischer & I. Moleenar (eds.) *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag, 215–237.

Wright, B. D. & M. H. Stone 1979. *Best test design*. Chicago: MESA Press.