John H.A.L. DE JONG (Ed.)

STANDARDIZATION IN LANGUAGE TESTING

AILA REVIEW - REVUE DE L'AILA

7

(1990)

Table of contents:

Guest-editor's Preface

A central purpose of the AILA Review is the presentation of individual fields of applied linguistic in such a way as to make them accessible and usable for scholars and practitioners beyond the scientific community primarily concerned. It was felt that the best way to introduce the field of language testing was by presenting one of its major preoccupations and reporting on how this is being dealt with, theoretically and practically.

Most readers have undoubtedly had to personally endure some kind of language test at some point in their life. In this sense applied linguists are not different from the general public. They do differ, however, in that as teachers and as researchers they make use of language tests, to assess their students and to confirm their hypotheses. Indeed language testing is so much an everyday phenomenon, that it's easy to overlook its complexity, as one tends to drive a car without pondering on how it has been put together or what might happen when something goes wrong.

Before a single question is drawn up, before a single item is written, language testers have to define what it is they would measure. Once this theoretical postulate about how reality operates is formulated - what it means, e.g., to have a functional command of a particular language mode - the postulate has to be transformed into a variable, through a set of items, each of which is intended to replicate the construct. The next step is to choose a measurement model, for, if the variable exists in reality, confronting individuals with the items should reveal variation. The measurement model is a mathematical expression of the intended workings of the set of items. Then data are collected and theory is confronted with reality. If the data do not conform to the model, the formulation of the construct is proven inappropriate and a new theory is required. Language testing, therefore, constitutes the link between theory and practice, and language testers operate in the messy, but fascinating, middle of the field. Under the pressure of having to make their instruments - and make them work - and, realizing the consequences that a malfunctioning of their measures may have for the individual, language testers are very much aware that models are models. They experience the conceptual inadequacies of models in representing reality in their attempt to capture real life behavior.

It is ironic therefore, that language testers who must in this way validate the models, are sometimes viewed as simple number crunchers, their preoccupations as down-to-earth, and their interests unscholarly. Data collection and model testing require expertise and patience. The contempt for the quantitative phase is simplistic. One could see this reaction as a kind of intellectual petulance: when faced with the complexity of the matter, the present indeterminacy of our knowledge, and the require-

ments of data collection and model testing, one tends to be frustrated, to throw the entire undertaking against the wall, so to speak, and walk away.

In 1989 AILA prepared a report to UNESCO on Language Teaching in the Perspective of the Predictable Requirernents of the Twenty First Century. The following suggestions were made in relation to language testing:

(1) *The definition of objectives and content in foreign language learning.*
Efforts should be undertaken to develop a functional scale of proficiency allowing for a criterion referenced system of both assessment of learner achievement and evaluation of curricular effectiveness.

(2) *The development and use of internationally comparable language tests. Research should be stimulated to enable actual comparisons of standardized objective language tests and testing procedures.*

(3) *The development of international standards for language testing. A code should be worked out concerning design, construction, and appropriate use of language tests and other requirements relating to test procedures and test quality.*

The third suggestion is presently being discussed and several national and international groups of professional language testers are investigating the possibility of forming an international association that would see the agree-ment on standards as one of its major goals. Therefore, a set of standards for language testing has been proposed to the international community of language testers for discussion and revision. Agreement on common standards would, in the view of this editor, constitute the condition for a pro-fessional organization.

The other two suggestions mentioned above were taken as the theme of a colloquium organized by the Scientific Commission on Language Testing and Evaluation at the AILA 1990 World Congress in Greece. This issue of the AILA Review offers specially commissioned papers on this theme, some of which were presented in earlier versions at the colloquium. It provides examples of how the issue of standardization is currently being dealt with by language testers from all over world. One of the striking aspects of the work in this area is its international character, reflected by the co-operation of researchers from different countries in joint projects.

More often than not tests used in the context of education are norm referenced. This is true for tests of language and of mathematics, for classroom tests and national examination papers. In most cases no pro-cedures are foreseen to equate test forms administered on different occas-ions, in different years, or in different school types. It is therefore impossible to control the level of educational outcome and to monitor shifts or drifts in the average abilities of students over time. Discussions on improving the educational system remain at the level of speculations. In order to evaluate changes in the system and/or shifts in the population, curriculum indepen-dent scales of functional proficiency will have to be developed. It is in this, original, sense that we interpret the concept of criterion-referencing: allowing for the interpretation of an individual's achievement as a position on a conti-nuum of developing proficiency.

It is noteworthy, that outside the educational system internationally accepted certificates do exist. A driver's licence, for instance, obtained within any country, will allow the bearer to drive a car in all other countries, without further questioning of the validity of the document. This

general acceptance does not imply an equal curriculum or even a comparable experience for all drivers from all countries. Clearly, the experience of a driver who has taken driving lessons in Mexico City differs from the experience of a driver in rural Germany. Also the experience of a young driver from a mountainous region will differ from the experience of a newly certified driver from a flat country. The implicit assumption underlying the international acceptance of the driver's licence is that some set of minimum requirements is sufficient to be able to manoeuvre a motorized vehicle. The necessary adaptations of a driving style to diverging circumstances are expected to be easily acquired whenever the need arises.

Both theoretical and practical aspects of standardization in language testing are presented in this issue. Examples are provided of how research can be undertaken, and what the results may be. Internationally comparable measures will provide individuals with certificates that express their achievement according to internationally recognized and accepted standards, and will thus create more freedom and mobility to the individual and a better control of the educational goals and assessment procedures across the nations.

In the first contribution Peter Groot argues for agreement among researchers on explicit criteria for construction, administration, analysis and evaluation of language tests used both in research and in education. His proposal would lead to comparability of test and research results across various research and educational settings. Groot regrets that tests are often the closing entry to balance the experimental budgets. He illustrates this claim with a number of examples taken from recent research literature on issues such as, the similarity of first and second language acquisition and the accessibility of Universal Grammar in second language acquisition.

Fred Davidson and Lyle Bachman illustrate the issue of test comparability with a paper on the Cambridge-TOEFL Comparability Study, which was intended to examine the statistical and content comparability of two internationally influential EFL test batteries. A hopeful finding from their study is the high degree of similarity between the two test batteries. Apparently the initial premise on differences between U.K. and U.S. educational measurement cultures did not result in major differences in test content or functioning.
In the past decade several efforts have been made to provide fairly detailed descriptions of hypothesized stages of language proficiency. These descriptions aim to offer criterion-referenced scales of developing language proficiency. In this context, David Ingram discusses the Australian Second Language Proficiency Ratings, in which the behavioral descriptors indicate the sort of tasks learners can carry out and how they are carried out at levels of increasing proficiency. In his contribution Ingram deals with the development and validation of the scale, but also with more general issues, such as, parameters of change in proficiency and the issue of general language proficiency. Ingram wishes to distinguish proficiency from communicative competence, arguing that the latter depends on traits other than the ability to use language, e.g., intelligence, education, general knowledge, introversion or extroversion.

The study reported by De Jong and Oscarson was conducted to investigate the possibility to use data gathered on EFL tests in one country to predict item and test characteristics if the tests were to be used in another country. Their results indicate that cross-national cooperation in developing and monitoring procedures and instruments for standardized objective assessment in the domain of language proficiency is worthwhile and that a foreign language skill can be scaled across different language backgrounds and given different curricula. They conclude that in spite of differences, the two educational systems in their study seem to lead to comparable outcomes.

By contrast, Shohamy and Stansfield argue that tests should match different contexts and purposes of language learning. They describe the development and validation of a speaking test that is intended to be used in two different contexts, a foreign language context and a second language context. Parallel versions of the test were developed to match these different contexts. Their results indicate that a tape mediated oral proficiency interview offers an efficient alternative to the type of direct oral proficiency testing as described in the contribution by Ingram. From the favorable validation study results in each of the learning contexts they conclude that both versions are appropriate for learners in the specific milieu where the language is learned and used. Their study shows that modern technology allows to combine human and material resources on different continents and work cooperatively on international projects. In their case, electronic mailing facilitated the internationalization of a project that otherwise would have been only national in scope.

In the last contribution Alex Olde Kalter and Paul Vossen try to find a justification for the use of existing U.S.-made tests in new, European contexts. They compare the workings of a test under its regular administration as a university admission test and its administration for certification purposes. They found, e.g., the same relative contribution of subtest scores under both circumstances and conclude that the different uses of the test do not result in differences in behavior of the candidates tested under each of the two conditions in their study. John Read in a discussion of their paper pleads for what he calls, more innovative approa-ches to the testing of English, as can be found in recent British tests and thus, in a way, links up to the topics dealt with in the second contribution to this issue by Davidson and Bachman.

It is hoped that this selection of papers will help to reveal both the points of agreement as the controversies that exist today in the field of language testing. There may be discussion on all major content issues, but language testers do agree that as professionals they have major obligations with respect to fairness and that language testing calls for clear definition of aims, meticulous research, and accountable reporting.

As a guest-editor I wish to thank AILA for the opportunity to present the field of language testing to the members of AILA-affiliates worldwide. Thanks are also due to all contributors for their cooperation and for meeting the deadlines.

John H.A.L. DE JONG

LANGUAGE TESTING IN RESEARCH AND EDUCATION: THE NEED FOR STANDARDS

Peter J. M. Groot
University of Utrecht

1 Introduction

The theme of the present issue of the AILA Review is standardization in testing. One reason for selecting this theme was that if there were certain explicit criteria for -construction, administration, analysis and evaluation of tests that all parties concerned would share, this would lead to an increased exchangeability of tests and comparability of test results across various research and educational settings. And this, in its turn, would have a highly beneficial effect on the quality of much foreign language research and teaching. Thus, in a paper on the effects of practice on foreign language learning, Ellis (1988) observes (in explanation of the often contradictory conclusions arrived at in various studies) that the measures used for the dependent variable (language proficiency growth) differ so widely that the results are hardly comparable. There are many aspects to standardization in testing: it can refer to procedures and terminology for describing the test contents, what it intends to measure, its format, administrative procedures etc. The American Psychological Association has issued various publications to stimulate standardization of testing practices such as the "Standards" (1974) and the more recent "Code" (1988). This article focuses on two important qualities of tests viz., reliability and validity and standardization of procedures and techniques to secure and demonstrate them.

To give a valid, general definition of scientific work is next to impossible. The areas covered by scientists are too complex and varied. Differences in the object of study are closely related to differences in the way research questions are generated and formulated, in the data collection methods to be used and techniques to analyze them. Still, some criteria common to all scientific endeavour can be formulated and one is the importance to be attached to the quality of the measuring instruments used to gather data that are relevant for answering the research questions. In the exact sciences, optimal attention for this aspect of research is standard procedure: a Nobel prize for instrumental research is no exception and constitutes proof that in those disciplines it goes without saying that the

quality of the instruments used determines the quality of the data obtained and, consequently, of the conclusions based on them. A different picture emerges, however, when one looks at studies in the area of second or foreign language acquisition. Often, the tests are the closing entry to balance the experimental budget: insufficient time is devoted to the selection or development of a relevant measuring instrument or to checks on the quality of the resulting data. And yet for this area of research, too, the same law holds that the validity of the conclusions drawn by the experimenter is wholly dependent on the quality of the tests used and the data they yielded. One can only guess at the causes of this negligence. Part of the explanation may be the fact that the object of study (human beings and their cognitive attributes) is intrinsically more difficult to measure than is the case in the exact sciences: human beings lack constancy over time and show a large intra- and inter-subject variability in disposition, motivation, attitude etc. This intrinsic difficulty in measuring properties of human cognition requires greater sophistication in selecting methods and techniques to collect and analyze data and it is this particular kind of know-how that is traditionally neglected or lacking in the linguistically orientated training of most researchers in this area. Besides, this training is also deficient in another respect: it has always emphasized the study of language as a system of formal rules and not as a means of communication. And it is this latter aspect i.e. the functional use of language for communicative purposes (with all the performance variables labeled "irrelevant" within Chomsky's formal approach to language) that is the main concern of language acquisition studies. It is not surprising, therefore, that many researchers find it difficult to operationalize communicative proficiency with its many linguistic and pragmalinguistic components.

The above observations may partially explain the frequent neglect of the measurement aspect in language acquisition research. The fact remains that checks on the quality of tests used both in research and educational contexts are essential for valid conclusions and decisions whether they refer to the confirmation or rejection of a hypothesis, or to determining the cut-off point for pass-fail or placement decisions. The last example is mentioned because there is, of course, no essential difference between tests used in research contexts to gather data on the tenability of some theory and tests used in educational contexts to measure students' achievement. The use may be different but construction, analysis and validation are subject to the same quality requirements.

2 Reliability and validity

Establishing the quality of measurements consists of various stages, each related to a particular aspect. To begin with, there is the aspect of reliability i.e. the accuracy or stability over time with which a trait (attribute, skill, knowledge, insight, etc.) has been measured. Reliability concerns the question to what extent differences in scores between subjects reflect dif-

ferences in the attribute measured or whether they are rather the result of accidental factors within the test (badly constructed or not enough items etc.) or without (cheating, underachievement due to indisposition etc.). If accidental factors have unduly influenced the test results, administration of an equivalent test at the same time or of the same test at a different time would have yielded a different rank order of the subjects; in other words: the scores cannot be interpreted as an adequate reflection of a trait, let alone the intended trait. In this article I will not deal with the various factors that influence the reliability of a test (cf. Cronbach, 1961 for an extensive survey of sources of error variance) but restrict myself to observing that there are various aspects to the concept reliability (such as sample representativity and objectivity) and consequently various ways of determining whether test scores are reliable. What they all have in common and this distinguishes them from ways to establish validity - is that they compare the results of two tests (or all test halves possible, like in the K.R.-20 estimate) that measure the same trait via the same method. Comparison of tests that (purport to) measure the same trait through different methods forms part of the validation process which brings us to the second aspect of the quality of a test viz., its validity. After reliability, which is an internal quality, has been demonstrated - and only then, because unreliable and meaningless results make any further search for validity useless - the experimenter has to establish the external quality of her test i.e. its validity. The question to be addressed then is: do the test scores indeed reflect the trait or ability she intended to measure or, in succinct Latin: mensuratum mensurandum? To illustrate the difference between reliability and validity one might take as an example the perennial debate about the cloze technique in language testing. On the basis of data that point to a high internal consistency and/or objectivity, advocates claim that cloze tests are a valid measure of reading comprehension (some even claim of overall language proficiency) while adversaries consider it a (pseudo-)vocabulary test. The controversy is not about the reliability of the cloze test: both sides agree that it appears to accurately measure some trait. The disagreement is about how to interpret the test scores: whether it measures reading comprehension or vocabulary, in other words about which trait it measures.

Clearly, demonstrating the validity of a test is much more difficult than determining its reliability. The latter is a necessary but not sufficient condition for the quality of the test that can easily be established through data like correlations with equivalent tests or estimates thereof. Demonstrating the validity of a test, however, presupposes the availability of a validated measure of the relevant trait which in the context of language proficiency testing is just not there. This is not surprising because prerequisites for a valid measure of language proficiency are a thorough analysis of what constitutes language proficiency, how it is acquired, its components, how they interact, their relative weight, etc. In short, they are the same questions that researchers in this field are trying to answer. Since criterion referenced validation is hardly possible in the field of language testing, researchers will have to resort to demonstrating kinds of validity

for their tests that are not related to a criterion measure viz., content and construct validity. Content validation is the process of investigating whether the selection of tasks in the test constitutes a representative sample of the larger set ("the universe of tasks" or "performance domain") that one is interested in. It is closely related to the concept of reliability but differs from it in the sense that reliability is a quantitative characteristic that is demonstrated with empirical, statistical data, while content validity is much more a qualitative property.

In construct validation one validates a test not against a criterion or another test but against a theory; a theory or construct is developed to be used as a provisional explanation of the test scores. The next step is to determine whether the theory is tenable. This will depend on the extent to which the relationship between the test and a number of variables as predicted by the theory agrees with the relationship between the test and these variables as it is actually found in subsequent investigations. To give one example, in establishing the construct validity of a reading comprehension test one might postulate, that vocabulary is an important component of the reading ability as operationalized in the test, in fact more important than, say, syntax. The hypotheses to be derived from this construct would then be that (1) there will be a close relationship between reading and vocabulary (that could be demonstrated by a high correlation (>.50) between the reading test and a vocabulary test) and that (2) this relationship will be closer than that between reading and syntax (to be demonstrated by a lower correlation between the reading test and a test of syntax of < .50). If these predicted correlations are indeed found, they could then be interpreted as a first confirmation of this aspect of the hypothetical construct and thus as evidence for some construct validity of the test. The above sketch of the process of construct validation makes it clear that this approach is basically the same as the principle underlying empirical research in general, viz., rejecting or maintaining a theory y collecting data that either disconfirm or corroborate hypotheses derived from it. This process is often cyclical because as Fiske (1971) states: "Concepts guide empirical research and empirical findings alter concepts. This interaction is the essence of science." It will also be clear that it is a laborious and time consuming process which may explain why it is often neglected.

Before dealing with some examples that illustrate the importance of good tests in research and education one final remark should be made on the issue of the quality of tests as indicated by their reliability and validity. It would be a fallacy to think that these properties are only relevant for tests: they apply more generally to any procedure for the systematic collection of data or measuring attributes. Thus, in a review article on classroom research, Chaudron (1988) observes that one of the major problems in interpreting the results of much classroom research into the possible effects of formal instruction on foreign language learning is the lack of data on certain aspects of the observational procedures and instruments used such as intra- and inter-rater consistency and the lack of an adequate

analysis of what has been observed. It will be clear that these aspects refer directly to their reliability and validity.

3 Examples

In the remaining part of this article I will give some examples of the crucial role of tests and the various ways in which they can fail, invalidating the conclusions based on the results they produced. The first two are taken from research into similarities and differences between first and second language acquisition. The third example concerns the educational use of tests.

3.1 Similarity of first and second language acquisition

Ever since the strong claim about the predictive power of a contrastive analysis of source and target language was abandoned, there has been a lively debate about the degree of identity between first and second language acquisition. This controversy is not of theoretical interest only but has direct implications for the actual practice of foreign language teaching: a strong similarity between the way a child acquires its first language and later second language acquisition would render many of the efforts that go into foreign language teaching redundant; it would make superfluous such activities as a contrastive linguistic analysis as the basis for selection and ordering of the input material, systematic correction of mistakes, gradation of the grammar etc.. Such proposals have indeed been put forward by, amongst others, Krashen (1982), Dulay, Burt, and Krashen (1982). The basis for these proposals were the results of a test called the "Bilingual Syntax Measure" (Burt, Dulay, & Hernandez-Chavez, 1975). This test (the BSM from now on) is meant to place children whose first language is Spanish or English on a scale of five levels of second language proficiency by eliciting certain morphemes (such as plural -s, 3rd person singular -s, continuous tense suffix -ing) and some simple verb forms. It is also claimed that, in the process, the test measures stages in language acquisition. Since the test results were indicative of a similar order of acquisition in the first and second language of their testees, they were interpreted by the authors (and many others since) as clear evidence for a strong similarity between first and second language acquisition in general and led them to put forward the teaching proposals mentioned above.

Their conclusions, however, give rise to many questions. For example, do the BSM data, which are limited to some areas of syntax only, warrant generalization to language proficiency in general, that is to say, including vocabulary, phonology, word order etc.? And, because the BSM testees were children who had acquired the second language in a natural environment, one must ask to what extent one can extrapolate this situation to a much more artificial context in which adults learn a foreign language by means of formal instruction? But apart from these questions concerning the far-reaching conclusions drawn by the authors, the test

data themselves are also open to serious criticism. 1 will only deal with some of the more obvious critical notes that can be leveled against the test. More thorough analyses containing explicit information on test contents, pro-cedures, scoring etc. – leading to the same conclusions as mine, incidentally – have been carried out by Rosansky (1979) and Cziko (1987).

As to the test's reliability, the authors, in the accompanying technical handbook, report admittedly low test-retest and inter-rater correlations. Their explanation is that their subjects - children - are not constant as to their language proficiency and, consequently, cannot yield stable scores. This may be so but a more obvious explanation is the small size of the sample (18 items) and the absence of an objectifying technique such as recording the subjects' test performance for later reference. Another point of criticism is that the reliability data reported are not based on the actual test scores but on their conversion to the 5 point proficiency scale, which may have distorted the actual state of affairs. Finally, statistical formulae for agreement have been used (like the kappa coefficient) that are meant for nominal instead of ordinal data. Taking all this into account it must be concluded that the BSM scores are not adequately reliable and that differences between subjects are to a large extent accidental and only partially the result of differences in some trait or other.

As to its content validity, the BSM claims to measure general syntactic skills. Its contents, however, only constitute a small-sized sample from the domain of syntax. Many other syntactic areas such as the passive voice, interrogatives, clauses etc. were not elicited. If a close relationship had been demonstrated between mastery of the latter structures and of the syntactic elements measured by the test, this would not necessarily have invalidated the BSM. Such a relationship is not mentioned anywhere.

The evidence offered for the construct validity of the BSM is also subject to serious criticism. As explained above, this kind of validity is demonstrated by finding confirmatory evidence for a theory (or construct) about what the test has measured: evidence for a relationship between test performance and other variables, as predicted by the theory. The data reported in connection with the construct validity of the BSM, however, only show a relationship between the construct that the authors accept as a provisional explanation of the test scores -i.e. the hypothesized parallelism between first and second language in the order of acquisition – and more or less identical measures of the same construct viz., earlier experimental versions of the BSM. It will be clear that this line of reasoning is circuitous, the more so because any external validity data are lacking.

The conclusion to be derived from the above analysis is that the BSM test results do not form an adequate reflection of the syntactic skills of the testees, let alone of their general language proficiency, and that the assignment to proficiency levels is dubious. The hypothesized similarity of first and second language as to the order of acquisition of certain syntactic elements cannot be regarded as sufficiently supported by the BSM data. Without taking up a position in this debate one can safely state that it is somewhat surprising to find that claims concerning parallelism between first and second language acquisition in general have been based

for so long mainly on the BSM data. A closer analysis of their quality does not sufficiently warrant this.

3.2 UG accessibility in second language acquisition

Another example of the vital importance in experimental research of the measuring instruments used and, consequently, of their standardization is taken from a field of research that has recently begun to attract attention, viz., universal grammar accessibility in second language acquisition. Extrapolating the well-known assumptions and hypotheses concerning the learnability of L1 as put forward notably by Chomsky, it is claimed by certain researchers in this area that the UG constraints on possible grammars that operate in L1 acquisition, are also unconsciously adhered to by L2 learners, both in natural and formal instructional contexts, irrespective of their first language and level of proficiency. This results in "knowledge" that enables learners of any (natural) second language to distinguish between what is grammatically possible and impossible. This theoretical stance has certain implications for the kind of errors that one expects learners to make (or, rather, not to make) and the areas that one expects to be impervious to LI transfer. It will be clear that this claim is worth while investigating, albeit less on practical than theoretical grounds. Even if UG-based strategies are followed in L2 acquisition, this is of relative importance only since the burden of the task facing the L2 learner lies elsewhere, namely in the acquisition of areas of the target language unaffected by these strategies such as its vocabulary, morphology and most of its phonology and word order. But apart from the hope for possible applications, however limited in scope, in L2 teaching there is a certain attractiveness to the idea of empirical verification of hypotheses derived from an LI acquisition theory through experimental data from L2 acquisition studies. These data might provide more insight into issues of common interest such as whether persisting L1 interference errors can be seen as evidence for a hierarchy in the learnability of certain L2 parameters and, if so, whether there is any connection between this hierarchy and a natural order of acquisition, as claimed by Pienemann (1984) in his Teachability Hypothesis. It will be equally clear, however, that for many reasons the above claim about UG-based strategies in second language acquisition is also a very difficult claim to investigate. In the first place, L1 acquisition is of necessity different from any later L2 acquisition since it is part of a much more general process of natural growth comprising simultaneous cognitive, motor and affective development. Consequently, the L2 learner is by definition different from the L1 learner and it does not seem unrealistic to hypothesize that some of these differences will entail differences in the L2 acquisition process. Thus, the difference in cognitive maturity, combined with the fact that there is already an L1, will directly influence the role played by metalinguistic reflection and awareness in the L2 acquisition process. It goes without saying that this effect will be even stronger in a situation where L2 learners are asked for explicit

statements on the grammaticality of sentences, which is the current method of collecting data in this field.

Experimental results reported by White (1987, quoted by Jordens 1989; 1990) may illustrate this possible artefact. She asked her subjects, Dutch E.F.L. learners, for grammaticality judgments on sentences in which the so-called "that trace effect" plays a role, such as

*Who do you think that saw Mary?
What do you suppose that Mary will do?

The difference between Dutch and English in this respect is that in Dutch extraction from both subject and object position is possible and in English only extraction from object position. The latter is considered by White to be a UG principle, Dutch being the exception. The grammatical sentences with object extraction were considered to be correct by her subjects more often than the ungrammatical sentences with subject extraction. White ascribes this result to UG "knowledge" since her subjects could not have derived their knowledge about the difference between what is correct and incorrect in English from their mother tongue and since it is highly unlikely that they were taught that English does not allow sentences with subject extraction. Jordens, however, rather than interpreting this result as a corroboration of the hypothesized role of UG in second language acquisition, offers an alternative, more plausible and imaginative explanation of the same data. Referring to unambiguous experimental evidence for the (object extraction) interpretation strongly preferred by Dutch native speakers in the case of similar Dutch sentences where both interpretations are possible, he ascribes the judgments of White's subjects not to UG based intuitions but to an analysis that finally turns out to be L1 based and metalinguistic.

Secondly, from a descriptive perspective, there does not always appear to be unanimity on what syntactic areas should be considered universal properties of natural languages (in other words, to belong to UG) and what areas are specific to particular languages. The interpretations of terms like parameters and markedness seem to differ across researchers. This means that certain experimental results that contradict the use of UG strategies are interpreted by advocates of UG accessibility in L2 acquisition as irrelevant to the issue and vice versa. To some, this may seem a definitory problem of minor importance, as it may be in studies addressing the logical problem of L1 acquisition. It is a major problem in empirical research, however, where one cannot hope to measure a trait or psychological property with any degree of accuracy without a satisfactory operational definition, i.e. a definition that states what operations subjects are asked to perform on what kind of material (in this case linguistic material). In fact, the problem is related to the more general methodological problem that arises when one borrows theoretical concepts concerning L1 competence from a non-experimental paradigm taking the ideal speaker listener as the starting point and attempts to operationalize these concepts in controlled, experimental investigations of L2 performance. In other

words, explaining the logical problem of acquisition with a theory that makes use of such concepts as insufficient positive evidence and the lack of direct negative evidence is one thing, trying to find confirmatory empirical evidence elicited from erratic L2 learners, whose linguistic behaviour is subject to all kinds of 'Irrelevant performance variables", quite another.

Thirdly, much of the linguistic stimulus material used in experimental studies of UG accessibility in L2 acquisition is semantically so intransparent that it is not unlikely that what test performance data actually reflect is not just (or not at all?) the subject's grammatical competence. The correct sentences below may serve as an illustration. They are taken from an experiment carried out by Bley-Vroman, Felix, and Ioup (1988). The first sentence was frequently rejected by non-native subjects:

What did John think Carol wanted her mother to give to the postman?

The second example is a correct sentence that was often rejected by native speakers on grounds admittedly unknown to the experimenters:

Which information would it be possible for Mary to persuade
Susan to tell the reporters?

This lack of semantic transparency may bring about noise elements in the measurement caused by the fact that normal language users (so: trained linguists excluded!) when confronted with linguistic material will show a strong tendency to want to understand it i.e. to assign meaning to the sentences. Anyone who has ever tried to teach L2 students the phonemes of a foreign language through nonsense words (so that students can better concentrate on the foreign sound) knows that this impossibility to "assign meaning" to the words frustrates students. Given this, it might well be the case that one cannot with impunity present sentences to subjects and ask them about the grammaticality of sentences they do not understand: in other words, to separate the purely syntactic from the semantic analysis. This implies that their judgments on the grammaticality will inevitably be influenced by the degree of semantic transparency of the sentences which, in its turn, will depend on their parsing difficulty. This phenomenon might also explain the often far from perfect scores of native speakers on the same stimulus material. This assumption concerning the "pre-emptive" link between the semantic and the grammatical aspects in verbal processing sounds even more plausible when one considers one of the claims of modern linguistic theory namely that certain universal properties shared by all natural languages (i.e. the UG principles and constraints) have mental correlates in the sense that they are the result of the way the human mind encodes and processes verbal information.

But however this may be, the issue of the role of semantic transparency is one which directly concerns the validity of the whole procedure of data collection in this field of research, and is as such worthy of

further analysis. One way of checking for this possible noise element in the test results would be to compare them to scores yielded by tests of the same attribute, utilising a different method, e.g., one that elicits spontaneous oral production data. I will come back to this technique of determining method versus trait variance at the end of this article.

The above observations on some of the methodological problems in this area of research are not so much meant to imply that researchers in this field are so naive that they are unaware of them. In fact, a heightened sensitivity to these issues seems to be developing (cf. Chaudron 1983, and Sorace 1988; 1990). They are much more meant to illustrate the problems in this kind of research and the consequent necessity of optimal attention for the one part of experimental work in this area that lends itself to systematic, objective checks and technical analyses of its qualities viz., the operationalization in tests of the hypothesised traits. Optimal attention for the construction, administration and analysis that is needed all the more in an area riddled with theoretical and methodological problems. To illustrate the point 1 am trying to make more specifically, 1 will select one aspect of tests that is often neglected viz., the instructions given to the subjects. In any testing situation it is, of course, important to carefully word the rubrics, to make clear to the testees what they are expected to do. But the instructions take on a crucial role in investigations where the experimenter is trying to tap a complex and subconscious trait (grammatical competence) through operationalizations (such as grammaticality judgments or intuitions) that are very often not adequately reliable and valid themselves: not with natives with their finished steady state grammars (so that there is no valid criterion measure), let alone with learners with their insecure, transitional grammars. If, next, one realizes that the test tasks almost inevitably appeal to metalinguistic awareness (which may vary widely across subjects but is not primarily the trait one is interested in) it is self-evident that the instructions play a key-role in setting testees on the same and right track. In this context, one major pitfall that is as obvious as it is difficult to avoid lies in subjects utilising their explicit grammatical knowledge in judging test sentences even when they are asked to judge not whether the sentences are grammatically correct but "possible" as was done in the experiment by Bley-Vroman et al. quoted above. But apart from the question whether this potential source of error variance can be neutralised by adequate instructions or not, the fact remains that they form an essential part of the testing procedure. In this context it is fortunate (although at the same time somewhat alarming that it took so long) that recently an explicit plea was put forward for standardization of this part of tests to be used in similar or replication studies in order to achieve a certain comparability of data from different experimental settings and thus create a common basis for explanations and conclusions (Bley-Vroman, Felix, and Ioup, 1988). Combined with more standardization in reports of experimental studies in this area of such data as individual, intra-subject and group, inter-subject consistency and rational equivalence of the test items, this will provide a solid basis for checks of the reliability of the results.

3.3 Differential justifiability

My last example to illustrate the importance of good tests is not taken from experimental L2 acquisition studies, but from the field of measuring educational achievement or, more generally, any context where tests are used for selective or placement purposes. It concerns the reliability of tests and a psychometric index derived from it viz., the standard error of measurement. As stated above, the reliability of a test indicates the accuracy with which a trait has been measured. If a test is not reliable the differences between the testees are accidental and the rank order found is not meaningful. Estimates of the various kinds of reliability indicate the degree of accuracy of the test's scores as a whole, while the s.e.m. is a measure of the accuracy of an individual score. It is derived from the reliability coefficient as follows:

$$Se = Sx \sqrt{(1 - r)},$$

where se is the standard error of measurement, Sx the standard deviation of the test scores, and r is the estimate of the test's reliability.

The following example illustrates the relationship between the standard error of measurement of a test and the justifiability of decisions based on the test scores. It concerns tests of foreign language listening comprehension at advanced level, such as are annually produced for nationwide use in the school system by the Dutch National Institute for Educational Measurement (Dutch acronym; CITO). These tests and the model underlying them have been thoroughly researched as to their theoretical and practical merits (Groot, 1975; De Jong and Glas, 1987). They are used as part of the final exams in foreign languages and the scores are converted to a 10 point scale used for awarding grade points. Their KR-20 reliabilities mostly lie between .75 and .85, which is satisfactory for tests of such a complex skill at an advanced level administered to the rather homogeneous Dutch school groups. The actual sample test to be dealt with here, is the 1990 form of the Test of Listening Comprehension of Spanish as a Foreign Language, which is psychometrically highly comparable to the listening tests for the other modern languages. It consisted of 50 items, the mean score was 35.55, the standard deviation 6.66, the KR-20 reliability coefficient .80, resulting in a standard error of 2.98. The pass-fail cut-off point was determined at 28.5, so that students with a score of 28 and lower failed. In order to see how the reliability of a test affects individual students, let us take the case of student A with a score of 29 and student B with a score of 28. The chances of student A being better at the trait measured than student B can be estimated on the basis of the standard error (which will be rounded off to 3 for ease of calculating): with a probability of 67% student A's "true score" (which is his/her hypothetical average score on a large number of equivalent tests) lies in between (29 - 3 =) 26 and (29 + 3 =) 32 and student B's true score in between (28 -3 =) 25 and (28 + 3 =) 31. As can be clearly seen, there is a considerable area of overlap,

which means there is a considerable chance that another equivalent test would have resulted in an inverted rank order for these two students. Still, on this test student B failed and student A passed. And this procedure, of course, affected many other students in the same way. Thus, tests with a low reliability and consequently a high standard error will place many students on the wrong side of the pass-fail cut-off: in fact, the higher the standard error the more students will be affected by an unwarranted pass-fail decision. If we take a hypothetical test with the same s.d. as the Spanish listening test above but a reliability of .64, the s.e.m. works out as 3.96. Student A's true score would then lie in between 25 and 33, and student B's between 24 and 32, resulting in an even larger area of overlap and the inherent higher chances of the observed rank order being false. Taking decisions that may affect students' careers on the basis of such a test hardly seems warranted. Only tests that possess adequate reliability should be used for that purpose, that is to say tests that possess what one might call "differential justifiability", i.e. the extent to which differences between students as registered by the test can be justified, explained, demonstrated to mean something.

4 Method versus trait variance

To conclude this plea for more concern for the quality of tests in research and educational contexts, I would like to draw attention to a simple but highly efficient technique to effectuate a first basic check on the validity of tests. It is clearly a way of establishing (a prerequisite for) validity rather than reliability since it is concerned with the influence of the method (to be understood in this context as referring to such format aspects as mode of presentation: oral/written, the nature of the questions: open/m.c., wording of the instructions, etc.) on the scores obtained and it is only useful when sufficient reliability has been demonstrated. It is a preliminary check on the validity of a test and should precede investigations into other kinds of validity such as content or construct (cf. Palmer, Groot, and Trosper, 1981). It is highly relevant in the context of what was said above about the crucial importance of the instructions in UG accessability experimentation and it is inspired by an investigation reported by Dekker (1987) into the effect of the instructions on the scores of a speaking test. In a differential treatment design he clearly demonstrates that the way the instructions are worded exerts a considerable influence on the scores and, consequently, on the reliability and validity. The technique I am referring to is called the "multitrait-multimethod" approach. In this approach one recognizes that a test score is a function of the trait (which is relevant signal) and of the method (which is irrelevant noise) and the experimenter tries to determine the relative contribution in the observed variance of method versus trait: in other words, to what extent the scores obtained are the result of the particular method used or of the trait measured. It will be clear that if the test scores have been unduly influenced by the method used this invalidates the test as a whole. The relative contributions of

method and trait in the test scores can be determined by comparing the results of two tests of the same trait utilising different methods, and of two tests that measure different traits but use the same method. If two tests that are meant to measure different traits but utilise the same method show a closer relationship (witness e.g., a higher correlation) than two tests that measure the same trait using different methods, there is evidence that the method used has exerted more influence on the test scores than the trait measured, which is clear evidence for insufficient validity of at least one of each pair of tests. The correlation matrix below illustrates this aspect of validity for a hypothetical set of tests.

	Trait x method 1 (e.g., open questions)	*Trait x method 2 (e.g., mc. questions)*	*Trait y method 1*	*Trait y method 2*
Trait x method 1	*1.00*			
Trait x method 2	*.72*	*1.00*		
Trait y method 1	*.40*	*.25*	*1.00*	
Trait y method 2	*.30*	*.35*	*.68*	*1.00*

The matrix consistently shows higher correlations between the tests measuring the same trait, irrespective of difference in method, than the tests that measure different traits, even if they use the same method. The results can be interpreted as evidence for a certain basic validity of the tests. They demonstrate that the experimenter has a certain control over the traits he wants to test, since measures operation arising the same (intended) trait in different ways converge, while measures operationalizing different traits in similar ways do not, or do so to a much lesser extent. If, however, method variance exceeds trait variance, that is to say if tests of different traits using the same method correlate higher than tests of the same trait using different methods, the experimenter will have to investigate possible causes. The cause can be simple and easy to remedy as when subjects underachieve due to lack of experience with the method (lack of "testpertise", one might call it). But it can also be more complex and give rise to fundamental questions and rethinking, for instance in a case where a trait appears to be so elusive that it cannot be measured independently of a method or where traits that were deemed different

turn out to be the same. In order to establish this kind of preliminary validity the research design or educational testing programme should allow for the testing of each trait by at least two distinct methods. If this is combined with utilising each method for the testing of at least two traits, the results will be data on the relative contributions in the test scores of method and trait, which is of particular relevance for convergent and divergent construct validation (cf. Campbell & Fiske, 1959). Clearly, integrating the above methodological precautions in the testing programme, whether it is used in a research or educational context, clearly takes extra time for construction, administration, data analysis and evaluation. But however time-consuming this may be, it is time well spent, since it forces researchers to take another, more sophisticated look at what they set out to measure, thus deepening their insight into the research or educational measurement problem and cautioning them against foolishly rushing in where angelic treading is called for. Determining the quality of tests is not always easy. The process of establishing construct validity, more often than not the only empirical option in language acquisition research, is an especially laborious endeavour, and the temptation to neglect this essential part of research is strong. 1 hope to have demonstrated that a fruitful and responsible use of test results is possible only when researchers and educators alike resist this temptation.

References

American Psychological Association (1974). *Standards for Psychological and Educational Tests*. Washington, DC: Author.

American Psychological Association. (1988). *Code for Fair Testing Practices in Education*. Washington, DC: Author

Bley-Vroman, R.W., SW. Felix & G.L. loup (1988). The accessibility of universal grammar in adult language learning. *Second Language Research*, 4,1.

Burt, M.K., H.L. Dulay & E. Hernandez-Chavez (1975). *Bilingual Syntax Measure*. New York: Harcourt Brace.

Campbell, D.T. & D.W. Fiske (1959). Convergent and discriminant validation by the multi-trait, multi-method matrix. *Psychological Bulletin*,56,2.

Chaudron,C. (1983). Research on metalinguistic judgments: a review of theory, methods, and results. *Language Learning*, 23, 343-377.

Chaudron, C. (1988). Classroom research: Recent methods and research findings. *AILA Review*, 5, 10-19.

Cronbach, L.J. (1961). *Essentials of Psychological Testing*. London: Harper Row Ltd.

Cziko, C. Bilingual Syntax Measure 1 (1987). In: J.C. Alderson, K.J. Krahnke, and C.W. Stansford (eds). *Reviews of English Language Proficiency Tests*. Washington, DC: TESOL.

de Jong, J.H.A.L. & C.A.W. Glas (1987). Validation of listening comprehension tests using item response theory. *Language Testing*, 4, 170-194.

Dekker, J. (1987). Het meten van spreekvaardigheid in een vreemde taal [Measuring speaking proficiency in a foreign language]. *Toegepaste Taalwetenschap in Artikelen*, 29, 3.

Dulay H.L., M.K. Burt, & S. Krashen (1982). *Language Two*. Oxford: Oxford University Press.

Ellis, R. (1988) The role of practice in classroom language learning. *AILA Review*, 5, 20-39.

Fiske, D.W. (1971) *Measuring the Concept of Personality*. Chicago, IL: Aldine Publishing Co.

Groot, P.J.M. (1975) Testing communicative competence in listening comprehension. In: R.L. Jones and B. Spolsky (eds) *Testing Language Proficiency*. Arlington, VA: Centre fort Applied Lingusitics.

Jordens, P. (1989). *Talen kun je leren: tussen taaltheorie en praktijk* [Languages are learnable: between language theory and practice] Dordrecht: Foris Publications.

Jordens, P. (1990) Linguistics and second language learning. *Toegepaste Taalwetenschap in Artikelen*, 36,1.

Krashen, S. (1982). *Principles and Practice in Second Language Acquisition*. New York: Pergamon Press.

Palmer, A., P. Groot & G. Trosper (1981). *The Construct Validation of Tests of Communicative Competence*. Washington, DC: TESOL.

Pienemann, M. (1984). Psychological constraints on the teachability of languages. *Studies in Second Language Acquisition*, 6, 186-214.

Rosansky, E.J. (1979). A review of the Bilingual Syntax Measure. In: B. Spolsky (ed) *Some Major Tests. Advances in Language Testing* Series, 1. Arlington, VA: Centre for Applied Linguistics.

Sorace, A, (1988) Linguistic intuitions in interlanguage development: the problem of indeterminacy. In: J. Pankhurst, M. Sharwood Smith & P. van Buren (eds) *Learnability and Second Languages*. Dordrecht: Foris Publications.

Sorace, A. (1988) Indeterminacy in first and second languages: Theoretical and methodological issues. In: J.H.A.L. de Jong & D.K. Stevenson (eds) *Individualizing the Assessment of Language Abilities*. Clevedon: Multilingual Matters.

THE CAMBRIDGE-TOEFL COMPARABILITY STUDY: AN EXAMPLE OF THE CROSS-NATIONAL COMPARISON OF LANGUAGE TESTS

Fred Davidson
University of Illinois, Urbana-Champaign
Lyle Bachman
University of California, Los Angeles

1 Introduction

The Cambridge-TOEFL Comparability Study (CTCS) was commissioned by the University of Cambridge Local Examinations Syndicate (LTCLES) in 1988 with two goals. First, the CTCS intended to examine the statistical and content comparability of two internationally influential EFL tests: the UCLES EFL test batteries and the Test of English as a Foreign Language (TOEFL) and its companion measures, produced by Educational Testing Service (ETS) in The United States. Second, the CTCS intended to spark a long term program of research and test development in the area of foreign language proficiency and its assessment. This program would be informed by the very special understanding that can only arise from cross-national educational research.

While this paper will summarize findings of content and statistical comparability of the UCLES EFL test, the First Certificate in English WC1) and TOEFL batteries, the main focus here is on the second goal of the study. We pose the question: what has the CTCS taught us about international language testing research? We offer the CTCS as an example of cross-national language testing research -possibly even as a research model to be replicated. However, we also wish to place the CTCS findings in a larger framework, i.e., international cooperation in language testing. Thus we shall comment upon how the CTCS design could be altered for future international language testing research.

In order to set the stage for this examination of the second CTCS goal, it is useful to provide some general descriptions of educational measurement and language testing in the U.K. and the U.S. In so doing, we are cognizant of Alderson's (1990) caution that any broad U.K.-LT.S. educational testing comparison is in danger of oversimplification; we will return to this point later.

1.1 EFL Testing and Educational Measurement in the United Kingdom

In the UX, large-scale and national educational tests are developed by examination agencies or boards, often loosely associated with universities. UCLES is an example of this; it is a department of the University of Cambridge, but it is a semi-autonomous money-earning entity, with much the same relationship to the university as a university publisher might have.

U.K. examination boards are intricately involved in the development of curricula in the U.K. educational system. They participate with members of government, school administrators, and other concerned parties on a regular basis to examine curricula and educational policy. For example, the recent interest in educational reform via modular curricula has been informed by input from UCLES and other examination boards. This modular curriculum development implies modularized testing, and UCLES is quite involved with that as well. On the whole, UCLES is a very typical British educational testing agency. Its contributions are much broader than simply the preparation of tests.

But UCLES does produce examinations. And since its tests are the result of interaction with many other persons and agencies in British education, the concept of an exam is itself broader than that of a test. In U.K. education, an exam is better characterized as a course of study. It denotes both a given measurement event and the curriculum leading up to that event. In this scheme, tests are usually seen as certifications. For example, when a candidate takes the FCE, the score is reported as a certified level of achievement. The candidate may take that certificate to, for example, a potential employer as proof of 'having the FCE.

In summary, large-scale educational measurement in the U.K. is characterized by the crucial role of large, semi-centralized testing agencies which confer with members of the education community. Many exams, UCLES EFL measures included, are really courses of study bolstered by such centralized effort. The question of relevance to our discussion here is: do the UCLES EFL measures match the degree of centralized control found in other U.K. examinations?

The answer is yes and no. In general, EFL tests in the U.K. are developed as courses of study, which imply curricula, and which imply meetings with EFL teachers and administrators in the U.K. and overseas -a model very similar to content area testing for U.K. schools. However, this is only a loose similarity, and UCLES is again a good example. In general, the UCLES EFL tests do not have the same national curricular imperatives driving their construction as do, for example, changes in modular science and math teaching and testing[1].

[1] Alderson (1990) has also pointed out that the UCLES EFL battery does not necessarily reflect U.K. language tests in general. With this we also agree, but we note that the CTCS did not include U.K. tests other than the UCLES measures so we cannot address this issue in the depth it deserves. Later U.S.-U.K. language test projects should include multiple measures from both countries.

1.2 EFL Testing and Educational Measurement in the United States

In the U.S., large educational measurement agencies also exist, and ETS shall serve well as an exemplar. Unlike the U.K. examination boards, however, these agencies generally do not see their role as advisory in development of national curricular reform; indeed, there really is no such thing as a national curriculum in the U.S. Critics have charged that U.S. testing agencies do exert such influence in an uncontrolled and undesirable manner (see Evangelauf, 1990, for a summary of a recent commission report on this issue), although that is not an overt public role for U.S. testing agencies.

Historically, U.S. education and many social services are under decentralized, locally autonomous control at the level of individual states. This is a principle which imbues all civics in the U.S. – states' rights to govern matters that are locally important. The principle of local autonomy is repeated within states. A state government is likely to delegate a large part of curriculum and assessment authority to local school districts. The net result, in U.S. educational culture, is a largely decentralized system with heightened local responsibility over educational design, implementation and assessment.

Testing agencies have evolved a rather different role in such a climate, when compared to those in the U.K. In the U.S. the testing agency, and ETS is again an exemplar, is a clinical, 'objective' provider of presumably reliable information, that is useful for making placement, advancement and other evaluation decisions in the educational system. The information provided is devised and produced by a psychometric epistemology which itself is clinical and detached. Therefore, tests are developed with great attention to issues of sampling in order to be sensitive to local diversity – states and local districts within states (but note, again, Evangelauf, 1990). If the test publishers did not attempt to accommodate local diversity they would lose their market.

As we did with EFL exams developed in the U.K. we must again pose the question: what of international EFL tests developed in the U.S. but influential mostly overseas? Do they follow this clinical, detached, 'objective,' psychometric model? Again, as with the U.K. situation, the answer is yes and no. Since TOEFL's international influence vastly exceeds that of any other U.S. EFL test, it can be discussed as a clear exemplar.

The TOEFL battery can be thought of as the prototypical "psychometric-structuralist" test. It draws from structuralist tradition and its approach to language which, as Spolsky (1978) pointed out, is readily compatible with so called "discrete-point" psychometric tests. The core test, the multiple-choice (m/c) TOEFL, is a multi-item assessment of highly discrete language tasks. In that sense, the answer is 'yes', TOEFL does represent the U.S. measurement perspective. However, the rest of the TOEFL battery is somewhat different. The Test of Spoken English (TSE), while not a face-to-face interview, is a step away from discrete-point testing. So too is the Test of Written English (TWE). These latter two tests

challenge, to some extent, the psychometric tradition in that they include more integrative, global, and judgmental measures.

We must emphasize, however, that the non-m/c elements of the TOEFL battery are not as widely administered as the paper-and-pencil m/c TOEFL itself. The TSE is taken primarily by international students wishing to be teaching assistants in U.S. universities, and the TWE is still not given as often as the m/c test. Thus, when one speaks of 'The TOEFL, one is really speaking of the paper-and-pencil m/c test. And it is the TOEFL that is a particularly good representative of content area educational measurement in the U.S., probably representing this tradition better than the UCLES EFL measures represent the U.K. measurement tradition.

Finally, just as the UCLES EFL measures may not accurately reflect language tests all over the U.K. so too the TOEFL battery does not accurately represent language tests available within the U.S. However, again, the CTCS did not investigate U.S. EFL measures other than the TOEFL, so this matter is left for future investigation.

Both the UCLES and the ETS tests of EFL are clearly products of their distinctive educational measurement traditions. At the same time, the UCLES tests have characteristics on which they differ from the U.K. tradition. Finally, neither the UCLES nor the ETS EFL tests are particularly representative of the full range of EFL tests currently available in their representative countries. However, the goal of the CTCS was to compare the UCLES and TOEFL EFL batteries, not the entirety of U.K. and U.S. educational measurement or EFL testing traditions. At the same time, an understanding of both of these traditions and of the place of the UCLES and TOEFL measures in these traditions provides a broader basis for understanding the results of the CTCS.

1.3 A Working Hypothesis

Both the UCLES EFL tests and the TOEFL battery reflect some aspects of their respective measurement traditions. U.K tests are team products which reflect and inform national educational policy, while U.S. tests are clinical, detached, operate in an educational setting where no national curriculum exists, and once established, are difficult to change. We could thus hypothesize that the differences between these two measurement traditions would be reflected in the tests themselves. Specifically, the TOEFL battery and the FCE should appear to measure different aspects of language ability, and these differences in test content should produce differences in test performance. If, however, actual comparison of the content of and results from the two tests reveals that they appear to measure the same aspects of language proficiency, then it may be that the direction in which each differs from its respective tradition brings the two tests closer together. If, however, actual comparison of data from the two tests reveals that they appear to measure the same aspects of language proficiency, then it may be that the direction in which each differs from its respective tradition brings the two tests closer together. If that is the finding

then perhaps something else may drive the tests toward a common ground.

We have operationalized this working hypothesis in two ways. First, we would expect clear and striking differences in expert judgments of test content between the TOEFL battery and the FCE. Second, we would expect clear and striking differences between examinee performance as well – particularly in the nature of the traits measured.

In order to achieve maximum generalizability of our findings on this hypothesis, we had to address a related research question: did the CTCS candidate sample accurately represent "typical" TOEFL and UCLES EFL test-takers? If so, then the findings of this paper speak to a broader audience, to precisely the focus stated above: international comparisons of language tests.

2 Method

This section describes the CTCS subject selection procedures, test instruments and administration, and the scoring and analytic procedures. Two principles drove the design of the CTCS:

(1)obtain a representative sample of both "typical FCE takers" and "typical TOEFL takers," and

(2)follow as closely as possible the operational administrative procedures of the two test batteries

By incorporating these two principles into the design of the study, we felt we could assure the greatest generalizability of our results.

2.1. Subject Selection

In order to include both "typical FCE takers" and "typical TOEFL takers" in our sample, we used existing UCLES records to identify world sites where normal FCE enrollments were quite high. We also identified sites with low Cambridge enrollments but high TOEFL enrollments based upon published information available from ETS (ETS, 1987). In general we classified half of our sites as 'Cambridge-dominant,' and half as TOEFL dominant,' intending thereby to balance the subject sample. Additionally, we wanted to assure that different geographic areas and native languages would be represented. After studying the population distributions of the FCE and the TOEFL, we settled on sites in the Far East, the Middle East, Europe and South America. These regions were subdivided to accommodate representative samples of different languages: Chinese, Japanese and Thai in the Far East; Arabic in the Middle East; Spanish, French and German in Europe; Spanish in South America. Given these considerations, we identified the following sites: (1) 'Cambridge-dominant': Madrid, Sao Paulo, Toulouse and Zurich, and (2) 'TOEFL-dominant': Bangkok, Cairo, Hong Kong and Osaka.

At each site, a site coordinator was responsible for selecting subjects and for scheduling and administering the tests. In the Cambridge-domi-

nant sites, this person was the local examinations officer responsible for Cambridge tests, while at the TOEFL-dominant sites the site coordinators were individuals who were familiar with and had access to local "typical TOEFL takers" and also had experience with local TOEFL administrations. In addition, at all the TOEFL-dominant sites there was also a small Cambridge operation, and the local Cambridge officer also participated in the study. All local CTCS personnel worked together closely to identify potential subjects, either from the pool of known Cambridge candidates registered for the December 1988 examination (in the Cambridge-driven sites), or from possible TOEFL candidates (in the TOEFL-driven sites).

2.2 Instruments and test administration

UCLES produces several EFL exams, of which the CTCS included two: the FCE and the Certificate of Proficiency in English (CPE). A pilot study suggested that the CPE is at a higher level than the TOEFL, whereas the FCE was slightly more comparable to the TOEFL. Since it was not possible to administer the FCE, the CPE and the TOEFL to all candidates, the role of the CPE was lessened. We will only report the FCE and ETS results here. The FCE is divided into four parts, called 'papers':

-FCE Paper 1, entitled "Reading Comprehension," contains two parts. All items are 4-choice m/c. 25 items focus apparently on use or usage and ten are passage-based reading comprehension questions.

-FCE Paper 2, entitled "Composition," presents the candidates with five prompts, from which the candidate chooses two, writing 120180 words in response to each.

-FCE Paper 3, entitled "Use of English," presents items that appear to test lexicon, register and other elements of English usage.

-FCE Paper 4 "Listening entitled Comprehension," is a tape-recording-plus-booklet test. Candidates listen to several passages and respond to items on each passage.

-FCE Paper 5 is a face-to-face oral interview. It was administered with one examinee and one examiner in some instances and as a group interview with two examiners and up to three candidates in others.

The second principle in our research design was to follow operational administrative procedures for both test batteries. Thus, the FCE given in the CTCS was the actual December 1988 operational administration. However, it was not possible to synchronize this with an operational international TOEFL administration, and we therefore used the institutional forms of the TOEFL and Test of Spoken English. By 'institutional,' ETS denotes a form of the test that is no longer used operationally as an international test. Since the institutional tests are "retired" forms of the international tests, ETS guarantees the content and statistical equivalency of the international and institutional forms. Institutional tests can be leased by a local institution for its own testing purpose. At the time of the CTCS there was no institutional writing test, since the Test of Written English was still classified as experimental by ETS.

The institutional TOEFL is organized into three sections:
-Section 1 - Listening Comprehension,
-Section 2 - Structure and Written Expression, and
-Section 3 - Vocabulary and reading comprehension.
All items are in a four-choice multiple choice format, though actual item types do vary. This format is identical to the operational international TOEFL.

The institutional version of the Test of Spoken English(TSE) is called the Speaking Proficiency English Assessment Kit (SPEAK). It includes both training materials and test materials for administering and scoring the test. Both the TSE and SPEAK are tape-mediated exams: the candidate listens to prompts on one tape recorder or on a master language lab recorder, looks at written and graphic prompt material in a booklet, and records responses onto
a second tape recorder.

Since there is no institutional counterpart for the Test of Written English (TWE), we consulted with experienced TWE examiners and produced a 'clone' called the Test of English Writing (TEW). The TEW had a structure exactly like the TWE: the candidates were given a printed sheet with a prompt containing verbal and graphic information, and were then given 30 minutes to write an essay based on that information on a separate sheet of paper.

All CTCS subjects took both test batteries: FCE Papers 1 through 5, TOEFL, SPEAK and TEW. In addition, subjects completed an extensive background questionnaire, which yielded demographic information on the CTCS sample group.

2.3 Scoring Procedures

A critical element in rigorously following operational procedures was the scoring of the tests. All the FCE papers were marked, or scored, according to standard operational UCLES procedures. Paper 1 was scored by optical scanner. Paper 3 was scored by examiners themselves using an examiner prepared key, while Paper 4 was scored clerically using predetermined examiner-prepared keys. Papers 2 (composition) and 5 (oral interview) were compositions rated subjectively by experienced examiners. Each of the two compositions for Paper 2 was rated on a six-band scale with a maximum of 20 points, which are then summed to a maximum of 40 points. The Paper 5 oral interviews were rated locally at each site by the examiners who administered the tests. The Paper 5 score of each candidate was a sum of ratings in the following component skill areas: fluency, grammatical accuracy, pronunciation of sentences, pronunciation of individual sounds, interactive communication, and vocabulary resource, each of which was rated on a six-point scale.

The TOEFL answer sheets were optically scanned at the University of Illinois, using an answer key provided by ETS. Each SPEAK was recorded on a cassette tape as a part of the administration, and these tapes were rated by experienced SPEAK raters at Iowa State University and the University of Illinois. Each SPEAK tape was rated by at least two raters, ac-

cording to standard SPEAK/TSE procedures, on the following scales: grammar, pronunciation, fluency and comprehensibility. The ratings of different raters were then averaged to yield a score for each subject. Each of the TEW scripts was rated by at least two experienced TWE readers at the University of Toronto, using a single 6-point scale, again, according to standard TWE procedures. The two ratings for each script were averaged to the nearest whole number to obtain a score.

2.4. Analytical Methods

2.4.1 Expert Judgment Content Analyses

The content of the FCE and TOEFL was characterized by using rating instruments called the "Communicative Language Ability" (CLA) and "Test Method Facet" (TMF) instruments. These instruments were based on the frameworks discussed by Bachman (1990).

The development of the CLA and TMF rating instruments was a cyclical process that involved trying out the instruments on a set of tests, analyzing data, discussing the results with the raters (applied linguists) and revising the instruments. Early versions of the instruments incorporated a system of counting occurrences of various linguistic and pragmatic categories, and the results of this approach in a pilot study have been reported by Bachman et al. (1988). These instruments were subsequently given to several applied linguists for evaluation and comment, and based on this evaluation, we decided to redevelop the instruments, basing them on ratings by expert judges. The redevelopment of the instruments had gone through three complete cycles by the time of the ratings reported here.

Throughout this process, rater agreement was checked by generalizability analyses using the computer program GENOVA (Crick and Brennan, 1983). Generalizability coefficients were regularly high, reflecting, perhaps, the high degree of communication between the raters as the instruments were refined. Early in the instrument development process the tests were judged by three raters, but as the instruments were refined, only two raters were involved since the generalizability analyses showed ratings to be consistent. In the data reported here, items from the structure and reading tests WE Papers 1 and 3 and TOEFL Sections 2 and 3) were rated by three expert judges, while items from the listening tests WE Paper 4 and TOEFL Section 1) were rated by two. All of the raters were experienced language teaching professionals, and two were also experienced language testing researchers.

The CLA and TMF instruments used in the CTCS divided ability and method into a number of components, called 'facets' (see Bachman, 1990). There were twelve CLA facets and 36 TMF facets. For each test, the rater was given a computer template file; the rater directly keyed in ratings on the CLA and TMF facets. These computer files were then edited to form datasets. Means and standard deviations across items and across raters were calculated using PC-SAS (SAS, Inc, 1988) and then compared across the test batteries.

2.4.2 Performance Data Analyses

All UCLES data files were prepared in Cambridge, while all ETS data files were prepared at the University of Illinois. All data merging and data file assembly was performed using SAS or PC-SAS (SAS, Inc: 1988). Descriptive statistics were computed with SPSS-X Version 3.0 (SPSS, 1988), while exploratory factor analyses were conducted using programs written for the IBM PC by John Carroll (Carroll, 1989).

Results

3.1 Description of Subject Sample

CTCS examinees reported demographic and background information on their TOEFL answer sheet and the background questionnaire. Most CTCS examinees were enrolled as students – 21.3% secondary level, 38% college level, and 17% in an English institute of some kind. The median age of CTCS examinees was 21 with a range of 14 to 58. 59.4% of the CTCS examinees were female.

Information of this kind is not routinely collected for the UCLES EFL exams, so no comparison can be made between the CTCS sample and the typical FCE population. Demographic information is available for TOEFL test-takers, however (Wilson: 1982,1987) and in general, the CTCS subjects matched those demographics well. Wilson reported a mean age of 21.4 years for individuals intending to apply to an undergraduate study program ("undergraduate degree planners") and 26.3 for graduate degree planners. This compares to a mean of 22.7 for the CTCS group. There is a slight difference in sex, however; Wilson reported that 72% of his group
was male, compared to 41% for the CTCS sample.

The representativeness of the CTCS sample in terms of test performance was also examined. Table 1 provides the observed means and standard deviations for all CTCS measures. Tables 2 and 3 provide comparisons of CTCS means and standard deviations for FCE and TOEFL, respectively, with their relevant norm groups. Published norms for the ETS measures are from the TOEFL Test and Score Manual (ETS, 1987), while the FCE norms are from Cambridge Examinations in English: Survey for 1988. (UCLES, 1988).

Table 1

Score Distributions, All Measures

VARIABLE	MEAN	STDDEV	MIN.	MAX.	N
FCE1	25.95	4.90	10	40	1359
FCE2	24.30	6.04	0	40	1357
FCE3	24.86	5.71	1	40	1353
FCE4	13.60	3.18	4	20	1344
FCE5	27.20	5.95	1	40	1381
TOEFL1	49.62	6.67	29	68	1448
TOEFL2	51.12	6.90	25	68	1448
TOEFL3	51.49	6.70	28	66	1448
TOEFL TOTAL	507.43	58.86	310	647	1448
TEW	3.93	.89	1	6	1398
SPEAK GRAM.	1.93	.45	0	3	1304
SPEAK PRON.	2.13	.38	0	3	1314
SPEAK FLUENCY	1.95	.44	0	3	1304
SPEAK COMP.	201.57	40.91	50	300	1304

Table 2

Differences between CTCS Group Means and Standard Deviations and UCLES Norms,December 1988

	N		MEAN		STD DEV	
Test	CTCS	UCLES Norm	CTCS	UCLES Norm	CTCS	UCLES Norm
FCEI	1,359	30,816	25.95	27.19	4.90	5.19
FCE2	1,357	30,818	24.30	26.07	6.04	5.22
FCE3	1,353	30,805	24.86	26.30	5.71	5.26
FCE4	1,344	30,936	13.60	14.47	3.18	3.25
FCE5	1,381	31,040	27.20	28.04	5.95	5.72

Table 3

Differences between CTCS Group Means and Standard Deviations and ETS Norms

	N		MEAN		ST DEV	
Test	CTCS	ETS Norm	CTCS	ETS Norm	CTC5	ETS Norm
TOEFL1	1,448	714,731	49.62	51.2	6.67	6.9
TOEFL2	1,448	714,731	51.12	51.3	6.90	7.7
TOEFL3	1,448	714,731	51.49	51.	6.70	7.3
TOEFL Tot	1,448	714,731	507.43	512	58.86	66
TWE/TEW	1,398	230,921	3.93	3.64	0.89	0.99
TSE/SPEAK						
- GRAM	1,304	3,500	1.93	2.43	0.45	0.39
- PRON	1,314	3,500	2.13	2.10	0.38	0.49
- FLCY	1,304	3,500	1.95	2.15	0.44	0.45
- COMP	1,304	3,500	201.57	221	40.91	.45

Notes to Table 3:
-ETS TOEFL data based on total group of examinees tested from July 1984 through June 1986 (ETS, 1987, Table 2, page 21);
-ETS TWE mean and standard deviation calculated from the frequency distribution given by ETS (1989, Table 5, page 16);
-ETS TSE data based on scores of 3500 persons tested in TSE administrations November 1981 through June 1986;
-ETS Norm values are as given by ETS; CTCS values are reported to two decimal places of precision.

Because of the large sample sizes, virtually all the differences between the sample and norm group means were statistically significant. It is thus more meaningful to discuss practical differences between the CTCS sample and the FCE and TOEFL norm groups. The largest difference between a CTCS mean and a published UCLES FCE mean is for Paper 2 (composition): 1.77 points, while the largest difference between a CTCS TOEFL mean and a published ETS TOEFL mean is on the TSE/SPEAK Comprehensibility rating: 19.43. In both cases the CTCS mean is slightly lower than that of the norm group. However, neither difference is very large when compared to the overall size of the scales (1.77 out of 40 and 19.43 out of 300) or to the norm group standard deviations (1.77/5.22 and 19.43/45).

In summary, we believe that the CTCS subjects' test performance is reasonably representative of their target populations, based on published normative data from UCLES and ETS. Furthermore, the CTCS subjects' demographic characteristics do not differ markedly from those typical of ETS test takers, though we cannot make such a claim about typical FCE takers due to lack of published demographic data. We thus feel that the similarities between the CTCS sample and target populations justify examination of the research hypotheses.

3.2 Reliability

Table 4 presents a comparison of observed CTCS reliability estimates with published reliabilities.

Table 4
Reliability Estimates

Test	k	N	CTCS	Norm
FCE1	40	1,394	.791	.901
FCE2		1,357	(Not available)	
FCE3	52	995	.847	.870
FCE4	27	759	.616	.705
FCE5		1,381	(Not available)	
TOEFL1	50	1,467	.889	.90
TOEFL2	38	1,467	.834	.86
TOEFL3	58	1,467	.874	.90
TEW		1,399	.896	.86

Notes on Table 4.

1. *CTCS and UCLES norm reliabilities are given to three decimals of precision. ETS reliabilities are given to two decimals in the published ETS reports. No k-sizes are given for any of the composition or speaking tests because they do not contain items.*
2. *CTCS reliabilities are calculated as follows:*
   -Weighted average coefficient alpha's for the CTCS sample are given across 2 forms of Paper 1 and across 3 forms of Paper 4.
   -Coefficient alpha is given for the CTCS sample for the single form tests: FCE3, and TOEFL1 through TOEFL3.
   -Generalizability coefficients are reported for the CTCS sample for TEW and SPEAK Comp(rehensibility) composite score.
3. *Norm reliabilities are based on the following samples:*
   -FCEl, for 164,256 examinees who took the FCE from December 1988 through December 1989, inclusive.
   -FCE3, for three random sample of 300 examinees each, from the June 1989 FCE administration.
   -FCE4, for random samples of approximately 300 examinees per form from the June 1989 FCE administration.
   -TOEFL, for examinees tested in the U. S. and Canada between December 1984 and February 1986 (ETS 1987).
   -TWE, norm reliability averaged via the Fisher Z-transformation across six administrations from July 1986 through May 1988 (ETS, 1989).
   -TSE, for 134 examinees tested in the fall of 1979 (Clark and Swinton, 1980; ETS 1982)
4. *The version of the Institutional TOEFL used in the CTCS contained 2 experimental/ non-scored items in TOEFL Sections 2 and 3.*

For the discrete item tests (FCE Papers 1, 3 and 4 and the TOEFL) classical internal consistency estimates were used. Due to normal variability in the UCLES exam administration procedure, two forms of FCE Paper 1 and ten forms of FCE Paper 4 were administered to CTCS subjects; reliabilities for those tests are averages across forms using Fisher's Z-transformation. Consistency of the averaged ratings for the TEW and SPEAK were estimated using single-facet generalizability studies with raters as the facet. Since FCE paper 2 and FCE Paper 5 do not include operational re-ratings, no reliabilities could be estimated for CTCS data on those two tests.

The FCE1 norm reliability is the KR-20 for test forms from December 1988 and December 1989, again averaged via the Fisher Z-transformation. FCE3 and FCE4 norms are Z-transformation averages across all test forms used in December 1989 only. ETS norms are given in the score interpretation manuals cited above.

With the exception of the TEW and SPEAK comprehensibility ratings, all reliabilities for the CTCS test scores were slightly below their respective norms. The observed CTCS reliabilities for the TOEFL battery and FCE3 are within normally acceptable limits, while those for FCE1 and FCE4 are below normally accepted test development standards.

3.3 Comparability of abilities measured

We would like now to examine the working hypothesis more closely. Specifically, to what extent are the differences between U.S. and U.K. examination cultures reflected in the content and performance data analyses of these two test batteries? Results of the content analysis are presented first, with exploratory factor analyses of performance data following.

3.3.1 Expert judgment content analysis

Tables 5 and 6 present a summary of meaningful differences in the CLA and TMF content analysis ratings. Mean ratings on CLA and TMF facets were computed across all test items for all raters for each test battery. The values in Tables 5 and 6 are mean facet differences that were greater than the standard deviation of the facet ratings of either test. They thus constitute those ability and method facet ratings which displayed the greatest differences across the two test batteries.

Table 5

Meaningful Differences on Communicative Language Ability Ratings

SUBTESTS	FACET	DIFFERENCE Magnitude	DIFFERENCE Direction
Listening			
(FCE4, TOEFL1)	Syntax	1.8	TOEFL > FCE
Structure			
(FCE3, TOEFL2)	Strat.Comp.	1.00	FCE > TOEFL
Vocabulary			
(FCEl, TOEFL3)	Lexicon	.99	TOEFL > FCE
Reading			
(FCEl, TOEFL3)	Lexicon	.63	TOEFL > FCE

Table 6

Meaningful Differences on Test Method Facet Ratings

SUBTESTS	FACET	DIFFERENCE Magnitude	DIFFERENCE Direction
Listening			
(FCE4, TOEFL1)	Relat. to Passage	.50*	TOEFL > FCE
	Genre: Unprepared	.86	FCE > TOEFL
	Cohesion: Lexical	.97	TOEFL > FCE
	Soc.Ling:Register	.72	TOEFL > FCE (+casual)
Structure			
(FCE3, TOEFL2)	Vocab.:Infreq.	.84	FCE > TOEFL (+frequent)
Vocabulary			
(FCE1, TOEFL3)	Vocab.:Infreq.	1.12	FCE > TOEFL (+frequent)
	Vocab.:Special.	.73	FCE > TOEFL (+general)
	Topic: Academic	.71	TOEFL > FCE
	Genre: Unprepared	.58	TOEFL > FCE (+familiar)
	Soc.Ling.:Register	.64	FCE > TOEFL (+casual)
Reading			
(FCE1, TOEFL3)	Vocab.: Infreq.	.57	FCE > TCEFL (+frequent)
	Context:American	.76	TOEFL > FCE
	Context: Academic	.78	TOEFL > FCE
	Cohesion:Reference	.94	FCE > TOEFL

** =slightly below the criterion for "meaningful" difference*

The 'meaningful' differences between the tests indicate that raters perceived the TOEFL as more complex syntactically, having less frequent vocabulary and a having tendency toward more academic topics and academic contextualization.

However, we must point out that these differences were observed in a minority of the total number of facet ratings. Twelve CLA facets and 36 TMF facets were amenable to calculation of means (some TMF facets are categorical and are not reported here), yielding 48 comparisons per test. Furthermore, four subscales were compared: Listening (TOEFL1 and FCE4), Structure (TOEFL2 and FCE3), Vocabulary (TOEFL3, part 1 and FCE1 part 1) and Reading (TOEFL3, part 2 and FCE1 part 2); this yields 192 separate comparisons of mean facet ratings. Tables 5 and 6, taken together, reveal that only 19 out of 192, or 9.9% of the comparisons were l meaningful' by the standard established above. By far the majority of TMF and CLA ratings were either (1) not salient – i.e., with extremely restricted range on both tests, possibly zero, or (2) similar across tests. Therefore, the most striking conclusion from these analyses is the overwhelming agreement by raters on the content similarity of the two test batteries.

3.3.2 Exploratory factor analysis

Three intercorrelation matrices were used for exploratory factor analysis (EFA): (1) correlations among the five FCE scaled scores, (2) correlations among the eight ETS standardized scores, and (3) correlations among thirteen measures: the five FCE variables and the eight ETS variables (see Appendix for matrices). For each matrix, standard procedures for determining the "best" factor model were followed. Initial principal axes were extracted with squared multiple correlations on the diagonals of each matrix. The eigenvalues from these initial extractions were examined for relative magnitude and graphed in a scree plot. As a further check on the number of factors to extract, the Montanelli-Humphreys parallel analysis technique was also used (Montanelli & Humphreys, 1976; Tucker, n.d.). Principal axes were extracted with the number of factors generally equal to one above and one below the number of factors indicated by the 'elbow' of the scree plot. These extractions were then rotated to both orthogonal and oblique solutions using Kaiser's (1958) normal varimax and Tucker and Finkbeiner's (1981) least-squares hyperplane fitting algorithm (DAPPFR), respectively. A final decision about the best number of factors to extract was made on the basis of simple structure (a pattern of near-zero and strong loadings) and interpretability (the groupings of salient loadings of the test scores on the rotated factors).

Tables 7 and 8 present the EFA results for the separate tests, while Table 9 presents the results for thirteen variables – the subtests of both batteries. In all three extractions, the oblique DAPPFR solution proved most interpretable, so for each a Schmid-Leiman transformation to orthogonal primary factors with a second-order general factor was performed (Schmid & Leiman, 1957). This permits the inter-factor correlations to be interpreted as a second order general factor, with the original oblique factors becoming first order orthogonal factors.

Table 7

Exploratory Factor Analysis of FCE Papers

VARIABLE	COMMUNALITY
FCEI	.54835
FCE2	.48888
FCE3	.62272
FCE4	.41468
FCE5	.32595

FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
1	3.18529	63.7	73.7
2	.63769	12.8	76.5
3	.48866	9.8	86.2
4	.41719	8.3	94.6
5	.27117	5.4	100.0

DAPPFR inter-factor correlation: .826

Orthogonolized Factor Matrix with Second-Order General Factor

	GENERAL FACTOR	FACTOR1	FACTOR2	h2
FCE1	.733	.275	.062	.617
FCE2	.689	.260	.057	.546
FCE3	.809	.433	.061	.846
FCE4	.679	.071	.241	.524
FCE5	.622	.024	.310	.484
Eigenvalue	2.515	.336	.165	3.016
Pct of h2	50.3	6.7	3.3	60.3

Table 8

Exploratory Factor Analysis of ETS Tests

VARIABLE	COMMUNALITY
TOEFL1	.53783
TOEFL2	.57999
TOEFL3	.57389
TEW	.37555
SPEAK GRAM.	.80099
SPEAK PRON.	.60725
SPEAK FLCY.	.77096
SPEAK COMP.	.89596

FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
1	4.93914	61.7	61.7
2	1.17112	14.6	76.4
3	.55103	6.9	83.3
4	.39869	5.0	88.2
5	.38118	4.8	93.0
6	.28032	3.5	96.5
7	.20493	2.6	99.1
8	.07357	.9	100.0

DAPPFR inter-factor correlation: .601

Orthogonolized Factor Matrix with Second-Order General Factor				
	GENERAL FACTOR	FACTOR1	FACTOR2	h2
TOEFL1	.654	.275	.258	.569
TOEFL2	.648	-.004	.532	.703
TOEFL3	.642	-.036	-.559	.726
TEW	.534	.094	.341	.410
SPEAK GRAM.	.668	.576	-.032	.779
SPEAK PRON.	.651	.404	.126	.603
SPEAK FLCY.	.694	.567	-.009	.791
SPEAK COMP.	.750	.650	-.038	.986
Eigenvalue	3.444	1.325	.798	5.567
Pct of h2	43.1	16.6	9.9	69.6

Table 9

Exploratory Factor Analysis of FCE Papers and ETS Tests

VARIABLE	COMMUNALITY
FCE1	.58958
FCE2	.52228
FCE3	.66459
FCE4	.48009
FCE5	.42786
TOEFL1	.59892
TOEFL2	.60101
TOEFL3	.61938
TEW	.39597
SPEAK GRAM.	.80465
SPEAK PRON.	.62949
SPEAK FLCY.	.77734
SPEAK COMP.	.89563

FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
1	7.48415	57.6	57.6
2	1.32523	10.2	67.8
3	.65258	5.0	72.8
4	.57734	4.4	77.2
5	.55311	4.3	81.5
6	.50183	3.9	85.3
7	.38833	3.0	88.3
8	.37212	2.9	91.2
9	.34312	2.6	93.8
10	.27587	2.1	96.0
11	.25262	1.9	97.9
12	.20044	1.5	99.4
13	.07325	.6	100.0

DAPPFR inter-factor correlations:				
	Factor 1	Factor 2	Factor 3	Factor 4
Factor 1	1.000			
Factor 2	.451	1.000		
Factor 3	.634	.751	1.000	
Factor 4	.743	.661	.824	1.000

Orthogonolized Factor Matrix with Second-Order General Factor

	GENERAL FACTOR	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	h2
FCE1	.754	-.031	.078	.175	.104	.617
FCE2	.711	.074	-.023	.270	.002	.584
FCE3	.820	-.026	.028	.341	-.004	.789
FCE4	.704	-.004	-.048	.053	.236	.556
FCE5	.621	.173	-.028	.015	.165	.443
TOEFL1	.776	.058	-.059	-.43	.284	.692
TOEFL2	.680	.049	.573	-.004	.010	.793
TOEFL3	.711	-.085	.419	.032	.102	.699
TEW	.581	.074	.223	.097	.012	.402
SPEAK GRAM.	.642	.621	.015	-.000	-.000	.798
SPEAK PRON.	.707	.333	-.026	.131	.030	.630
SPEAK FLCY.	.676	.555	-.021	.007	.045	.767
SPEAK COMP.	.705	.719	.040	.011	-.034	1.018
Eigenvalue	6.401	1.377	.570	.252	.189	8.789
Pct of h2	49.2	10.6	4.4	1.9	1.5	67.6

The results given in Tables 7, 8 and 9 corroborate the similarity of the two tests found in the content analysis. The within-test EFAs (Tables 7 and 8) reveal a large orthogonalized general factor; in the FCE it accounts for 50.3% of the variance of the model, while in the ETS measures it is 43.1%. The primary factors in each test display reasonably simple factor structures. For the FCE, these generally reveal a written mode vs. a spoken mode (FCE 1,2,3 vs. FCE 4 and 5, respectively). Among the ETS measures, the SPEAK defines a fairly clear factor, and the written mode m/c tests define another, with the TOEFL m/c listening test loading fairly evenly on these two factors. There is some similarity across the test batteries when viewing these within-test results in that the written mode tends to define a factor separate from the spoken mode.

Table 9 gives EFA results for the tests combined. The most striking feature of Table 9 is the size of the orthogonalized general factor (49.2% of model variance). Additionally, primary factor 1 is defined by the SPEAK and FCE5 – an oral factor; factor 2 is defined by the structure and reading/vocabulary sections of the TOEFL and by the TEW – an ETS written mode factor; factor 3 is defined by FCE1, 2, and 3 – an FCE written mode factor; and factor 4 is defined by the two listening tests.

In summary, we can state that there is a great deal of factor similarity across the test batteries, both when comparing the within-test EFAs and when considering the across-test combined EFA. The specific factors in Table 9 account for only 18.4% of the variance. They are salient but small, and they are outweighed by the strong inter-factor correlation expressed as the orthogonalized general factor. Another way to phrase the EFA results is to state that the performance data reveal a large common shared language trait measured across the two batteries. However, since the factor structure exhibited by any given set of test scores will reflect both the abilities of the specific groups tested and the characteristics of the particular tests used, the general factor found in our analyses does not necessarily correspond to the language g-factor that has been found in

earlier studies (e.g. Oller, 1979; Carroll, 1983; Bachman and Palmer, 1982), and which has been the subject of much historical debate. Rather, it is most appropriately interpreted as a general trait shared among the CTCS subjects, as measured by the FCE, TOEFL, SPEAK and TEW.

4 Discussion

Similarities were found in both expert judgments of content and in the factor structures of these two test batteries. Why might this be so? The initial premise of this paper was that differences between the U.K. and U.S. educational measurement cultures and language traditions that underlie the FCE and TOEFL would result in major differences in the tests themselves and in test performance.

We would like to offer several possible explanations for the degree of similarity discovered. First, it is probably the case that neither the FCE nor the TOEFL battery adequately represents the extreme case of either measurement culture. As the world is increasingly internationalized in educational trends, perhaps very large international tests tend to evolve toward a common ground.

A second explanation might be sought in our sampling design. If the operational TOEFL and FCE populations are in fact actually different, then our sampling design may not have adequately tapped that difference. However, this reasoning does not seem to hold, since the performance of the CTCS sample was very similar to published norms for both the FCE and the TOEFL. Thus, a more likely explanation would appear to be that on the whole, the FCE and TOEFL populations are more similar than we expected.

Finally, the differences between U.K. and U.S. educational measurement cultures may not be as great as we initially postulated. A more useful way of characterizing this difference might be that given by Vickers (1983): determination of what evidence is prioritized in a given measurement event. In U.K. educational measurement decisions, expert judgment and team-based test moderation do play a crucial role. Yet UCLES routinely calculates psychometric statistics in its item processing. Conversely, in the U.S., psychometric epistemology may be dominant in item development and decisions about candidates, but expert judgment is also involved – for example, in the assessment of item bias. What is critical is not what each measurement culture does, but in what it prioritizes. In reality, each culture may share many testing practices with the other, differing in the priority each practice receives.

In closing, we would like to offer the CTCS as a model for future cross-cultural comparisons, with the following cautions:

(1) Given the diversity of tests that is likely to characterize the educational and EFL measurement practice of any country, generalizations about system-wide differences cannot rest upon comparison of single test pairs. Just as we should not generalize that the FCE represents U.K. assessment epi-stemology nor the TOEFL that of the U.S., we should not conclude that our

findings generalize to all comparisons between U.K. and U.S. measurement. In future projects such as the CTCS, if one of the purposes is to make generalizations about culture-wide differences in measurement, it is extremely important to sample each measurement culture widely, comparing many test pairs from the full range of tests in countries involved. (2) There is likely to be an interaction between educational measurement differences across cultures and the operational populations of tests developed in those cultures. We would expect that tests produced and administered by institutions in different countries would be taken by different types of test takers internationally. If the comparison is made on the basis of tests given to very similar populations, then not much will be discovered about differences between the two measurement cultures. Therefore, in order for cross-cultural comparison studies to be sensitive to actual differences in operational populations, their sampling designs need to implement expected convergence and divergence in the two operational populations. That is, the sample should include subjects who could reasonably be expected to be very similar to each other, and also those who could be expected to be very different. In the case of the CTCS, this would have involved including both U. K. and U. S. college-bound students (the TOEFL-type person) and U. K. and U. S. non-college-bound persons (the FCE-type). If it turns out that the two operational populations are quite similar after all, this will also be clear from the results. (3) In any cross-cultural comparison study, it is also important to analyze not only test content and performance, but also the decision priorities which drive each assessment culture. Expected differences could be moderated by differences in prioritization such as we have cited above.

We realize that these recommendations suggest much larger and more complex international language testing projects, but we believe such projects are both possible and worthwhile. We have learned, often with memorable frustration, of the negotiated nature of cross-national studies (Schwille and Burstein, 1987). At the same time, we have learned that the give-and-take required in cross-national research can be accommodated. Furthermore, we have come to believe that if different measurement cultures are truly to learn from each other, such research is essential. On a personal level, it enables the individual researchers, through day-to-day collaboration in planning, design, implementation and reporting, to learn about the operational manifestations of each other's measurement traditions and to better understand those features that are particularly valued by the other tradition and why they are so valued. Such research is also important, we believe, since it has the potential, demonstrated by the CTCS, of revealing a great deal about not only the test batteries under investigation, but also about the measurement cultures of which they are a part, a great deal that we may not learn if we were to investigate either test battery or measurement culture in isolation.

Acknowledgements

We are grateful to the University of Illinois, the University of Cambridge Local Examinations Syndicate, and the University of California, Los Angeles for support with various phases of this research. We would also like to thank Inn-Chull Choi and Katherine Ryan for their work in the pre-paration of the test score data, and Brian Lynch for assistance in the preparation and analysis of Tables 5 and 6. Our other content analysis raters also deserve thanks: John Foulkes, Mark Harrison and Terry Santos.

References

Alderson, J. Charles (1990).
Comments on the Cambridge-TOEFL Comparability Study. Paper presented at the colloquium on the CTCS, TESOL, San Francisco, March 1990.

Bachman, Lyle F, (1990).
Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Bachman, Lyle F. and Fred Davidson (1990).
A comparison of the abilities measured by the Cambridge and Educational Testing Service EFL Test Batteries. *Issues in Applied Linguistics* 1,1: 31-57.

Bachman, Lyle F., Fred Davidson and Brian Lynch (1988).
Test method: the Educacontext for performance on language tests. Paper presented at the Annual Meeting of the American Association for Applied Linguistics, New Orleans, December 1988.

Bachman, Lyle F., F. Davidson, Katharine Ryan and Inn-Chull Choi (1989).
An investigation into the comparability of two tests of English as a foreign language: the Cambridge-TOEFL comparability study. Cam-bridge, UK: University of Cambridge Local Examinations Syndicate.

Bachman, Lyle F., A. Kunnan, Swathi Vanniarajan and Brian Lynch (1988).
Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language Testing* 5, 2: 128-59.

Bachman, Lyle F. and Adrian S. Palmer (1982).
The construct validation of some components of communicative proficiency. *TESOL Quarterly* 16,4:449-65.

Carroll, John B. (1983).
Psychometric theory and language testing. In: J. W. Oller Jr. (ed) *Issues in Language Testing Research.* Rowley, MA: Newbury House, pp. 80-105.

Carroll, John B. (1989).
Exploratory factor analysis programs for the IBM PC (and compatibles). Chapel Hill:Author.

Clark, John L. D. and Spencer Swinton (1980).
The test of spoken English as a measure of communicative ability in English-Medium instructional settings. TOEFL Research Report 7. Princeton: Educational Testing Service.

Crick, Joe E. and Robert L. Brennan (1983).
Manual for GENOVA: A Generalized Analysis of Variance System (ACT Technical Bulletin No. 43). Iowa City, IA: American College Testing Program.

Educational Testing Service (1982). *Test of Spoken English: Manual for Score Users*. Princeton, NJ: Author.

Educational Testing Service (1987). TOEFL Test and Score Manual, 1987-88 Edition. Princeton, NJ: Author.

Educational Testing Service (1989).
Test of Written English Guide. Princeton, NJ: Author.

Evangelauf, J. (1990).
Reliance on multiple-choice tests said to harm minorities and hinder reform: Panel seeks a new regulatory agency. *The Chronicle of Higher Education* 26:37, Al +A31.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23: 187-200.

Montanelli, Richard G., Jr. and Lloyd G. Humphreys (1976).
Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: a Monte Carlo study. *Psycho-metrika* 41:34148.

Oller, John W. Jr. (1979). *Language tests at school: a pragmatic approach*.
London: Longman.

SAS Institute, Inc. (1988). *SAS Guide to Personal Computers: Language. Version 6.* Cary, NC: Author.

Schmid, J. and J. M. Leiman (1957). The development of hierarchical factor solutions. *Psychometrika* 22:53-61.

Schwille, John and Leigh Burstein (1987).
The necessity of trade-offs and coalition building in cross-national research:a critique of Theisen, Achola and Boakari. *Comparative Educational Review 31*, 4:602611.

Spolsky, Bernard (1978). Introduction: linguistics and language testing. In B. Spolsky (Ed.) *Approaches to Language Testing*. Arlington, VA: Center for Applied Linguistics.

SPSS Incorporated (1988). *SPSS-X User's Guide*. 3rd Edition. Chicago:
Author.

Tucker, Ledyard R. (n.d.). Functional representation of Montanelli-Humphreys weights for judging the number of factors by the parallel analysis technique. Champaign: Author.

University of Cambridge Local Examinations Syndicate (UCLES) (1988).
Cambridge Examinations in English: Survey for 1988. Cambridge: Author.

Vickers, Geoffrey (1983). *The Art of Judgment*. New York: Harper and Row.

Wilson, Kenneth M. (1982). *A Comparative Analysis of TOEFL Examinee Characteristics, 1977-1979. TOEFL Research Report 11.* Princeton, NJ: Educational Testing Service.

Wilson, Kenneth M. (1987).
Patterns of Test Taking and Score Change for Examinees Who Repeat the Test of English as a Foreign Language. TOEFL Research Report 22. Princeton, NJ: Educational Testing Service.

APPENDICES

Intercorrelations among FCE Papers

	FCE1	FCE2	FCE3	FCE4	FCE5
FCE1	1.000				
FCE2	.581	1.000			
FCE3	.707	.668	1.000		
FCE4	.537	.494	.566	1.000	
FCE5	.465	.444	.474	.496	1.000

Intercorrelations among ETS Tests

	TFL1	TFL2	TFL3	TEW	SP-GR	SP-PR	SP-FL	SP-CO
TOEFL1	1.000							
TOEFL2	.552	1.000						
TOEFL3	.565	.716	1.000					
TEW	.448	.540	.517	1.000				
SPK-GRAM	.584	.437	.382	.389	1.000			
SPK-PRON	.588	.464	.474	.450	.642	1.000		
SPK-FLCY	.615	.419	.412	.434	.784	.675	1.000	
SPK-COMP	.640	.473	.443	.435	.893	.752	.873	1.000

Intercorrelations among ETS Tests and FCE Papers

	TFL1	TFL2	TFL3	TEW	SP-GR	SP-PR	SP-FL	SP-CO	FCE1	FCE2	FCE3	FCE4	FCE5	
TOEFL1	1.000													
TOEFL2	.554	1.000												
TOEFL3	.565	.718	1.000											
TEW	.439	.543	.518	1.000										
SPK-GRAM	.587	.437	.383	.387	1.000									
SPK-PRON	.586	.472	.476	.448	.641	1.000								
SPK-FLCY	.614	.423	.408	.429	.789	.668	1.000							
SPK-COMF	.642	.478	.447	.433	.893	.747	.874	1.000						
FCE1	.606	.570	.627	.441	.473	.524	.501	.534	1.000					
FCE2	.536	.518	.502	.470	.501	.567	.488	.538	.576	1.000				
FCE3	.587	.629	.638	.528	.491	.592	.522	.546	.700	.662	1.000			
FCE4	.611	.443	.497	.424	.497	.527	.524	.535	.530	.494	.563	1.000		
FCE5	.555	.410	.404	.366	.543	.531	.561	.572	.455	.452	474	.492	1.000	

THE AUSTRALIAN SECOND LANGUAGE PROFICIENCY RATINGS (ASLPR)

David E. Ingram
Griffith University, Australia

1 Introduction

The Australian Second Language Proficiency Ratings [Ingram and Wylie 1979/851 were first released in January 1979 and have undergone several reviews and revisions since then in the course of experience in the scale's trialling, use and the preparation of new versions (e.g., for languages other than English and for the assessment of special purpose proficiency). The ASLPR has now been in widespread use around Australia and abroad for about a decade and has contributed substantially to the development of other scales including the various sub-scales of the joint AustralianBritish test of English for overseas students, the International English Language Testing System (IELTS), and, it seems from the similarity of some descriptors, the later released ACTFL Guidelines.

The ASLPR is a proficiency rating scale containing nine defined proficiency levels in each of the four macroskills (0, 0+, 1-, 1, 1+, 2, 3, 4, 5) and a further three (2+, 3+, 4+) that are available for use but are undefined. The number system was adopted to be compatible in whole-number levels with the FSI Scale, which all editions of the ASLPR acknowledge as having been the starting-point for the development of the ASLPR. The names of the levels also bear similarities to those of the FSI levels (where they equate) though some have been modified to better fit the different nature of the ASLPR: Zero Proficiency (0), Initial Proficiency (0+), Elementary Proficiency (1-), Minimum Survival Proficiency (1), Survival Proficiency (1+), Minimum Social Proficiency (2), Minimum Vocational Proficiency (3), Vocational Proficiency (4), and Native-like Proficiency (5).

The ASLPR descriptors seek to describe observable language behaviour at nine levels as proficiency develops from zero to native-like. Each macroskill is described and rated independently in four separate but conceptually related scales so that a learner's proficiency is described and rated in a profile such as S:1+, L:2, R:1+, W:1.

The behavioural descriptions indicate the sorts of tasks learners can carry out at each proficiency level and how they are carried out, i.e., the nature of the language forms that appear. Each descriptor seeks to be as comprehensive as is reasonably possible indicating the tasks and the sort

of linguistic features (e.g. syntactic forms, lexis, cohesive features, phonology, functions, register flexibility, etc) that appear. These are not identified in terms of absolute items that are used but rather as a general description of how they appear; thus, for example, actual syntactic items are not listed but rather the overall shape of an utterance, the general features that have emerged, and the extent of the hierarchical development are indicated. Unlike the FSI Scale, the ASLPR does not assert that its levels are 11 absolute" or discrete but, rather, each descriptor exists in the context of the whole scale and in relation to adjacent descriptions. Hence, to show gradation, some descriptive features are unavoidably comparative in nature and omission of a feature at one level that is included at the next implies that it is non-existent or insignificant at the lower level. At the lower end of the scale, the levels cluster more closely because learners tend to progress through the lower levels more rapidly than the higher, greater differentiation is required because more students are in classes at this lower level, and the overt and more readily observable changes in language behaviour are more striking here.

The descriptors seek to provide an overall description of the learner's language behaviour at each level and, in rating, it is the observer's global assessment that is important. In this sense the descriptors are not checklists and the learner's proficiency is rated as that level whose description best matches the language behaviour observed. Although this interpretation of the descriptors may seem to make the assessment procedure more subjective than if each description were regarded as a checklist of essential features, it caters more adequately for the complexity of language behaviour and the fact that the different features within that complex whole may develop at different rates. Consequently, though one would generally expect some relationship between the different features (otherwise the notion of a rating scale and of a common schedule of development become meaningless), there might be relatively minor and possibly compensating variations within the total picture of a learner's behaviour at a particular proficiency level or there might be "maverick" development of some feature which has responded to cultural patterns or particular situational experience without necessarily reflecting on the learner's overall level of performance in that macroskill. Thus, for instance, in Speaking, a Vietnamese learner of English may reveal less development in phonology but a higher level of development in syntax; during the formal trials of the ASLPR, a Hong Kong businesswoman showed considerable ability to comprehend and perform operations with sums of money even though her overall Listening proficiency was not above L:0+.

The scale is presented in three columns. The first is the General Description column where a general description of the observable language behaviour is provided. This column is common to all versions of the ASLPR including the language-specific versions (for Japanese, French and Italian [Ingram and Wylie 1982, 1982a, 1982b) and the ASLPR for Special Purposes that is currently being developed. The second column, Examples of Specific Tasks, provides examples of the sort of language which

occurs at that level. The third or Comment column simply provides definitions of terms or other explanations to clarify something said in one of the other two columns.

2 Origins of the ASLPR

While seeking, in the course of his doctoral research in 1975, to identify the sort of foreign language skills learners characteristically display on matriculation and prior to entry to a tertiary language programme [Ingram 19781, the present writer became acutely conscious of the need for some means of objectively and comprehensibly stating learners' language proficiency. The opportunity to initially develop a proficiency rating instrument came in the context of developing new on-arrival programmes in English as a Second Language for the Australian Adult Migrant Education Programme and work commenced on the ASLPR in the latter part of 1978 with the initial version being released in early 1979. In that context, four needs led to the development of the scale: the need for some overall concept of language proficiency development as a framework within which a coherent series of courses could be planned, the need to emphasize the development of practical proficiency as the purpose of the programme, hence, the need to be able to state goals in unambiguous proficiency terms, and the need for assessment procedures also to emphasize the attainment of practical proficiency.

In developing the ASLPR, its authors drew on psycholinguistic studies of language development, their own intuitions and experience as teachers of several different languages (especially ESL, French, Italian and German), and the FSI Scale [Foreign Service Institute School of Language Studies 19681 to prepare tentative descriptions of language behaviour at the selected proficiency levels. Then, over a period of months and subsequently years, they elaborated, assessed, reviewed and revised these descriptions through face-to-face interviews with large numbers (effectively some hundreds) of adult and adolescent second and foreign language learners particularly of English but also of other languages (especially French, Italian and Japanese). In most cases, the focus of these interviews and hence the scale descriptions has been on general proficiency but more recently attention has been paid to specific purpose proficiency with the intention of producing specific purpose versions of the ASLPR. Formal trials of the ASLPR were conducted in Australia, China and the United States to assess its validity and reliability when used with adult and adolescent learners of ESL, adult and adolescent learners of French, Italian and Japanese as foreign languages, and when ratings are assigned by first and foreign language speakers of the target language. Some of the trial results are reported below and more fully in Ingram 1982.

The FSI Scale was chosen as the starting-point because, at that time, it was the most widely accepted and used and most thoroughly researched performance-based scale available. As an instrument suitable for the purposes for which the ASLPR was developed, however, it was rejected for

several reasons. First, it is strongly influenced by the fact that it was developed for use with well-educated civil, military and foreign service personnel learning languages other than English with the target variety being the "educated" dialect of the language whereas a scale was required that could be used with learners of a wide variety of educational backgrounds learning any of a variety of dialects. Second, the FSI Scale has insufficient differentiation at the lower end where changes in observable behaviour are most evident and most readily described, and where there is need to stream learners into a variety of classes. Third, FSI descriptions were available only for "speaking" (which included listening) and reading" whereas, if there is justification for using this type of proficiency measurement for speaking and reading, it is illogical to exclude writing and the intermingling of macroskills disregards the empirical evidence (of teachers as well as from subsequent data accumulated in using the ASLPR [e.g., Ingram 1982a]) that different macroskills may be at different proficiency levels. Fourth, it was felt that the descriptors themselves were not entirely adequate, that grammatical development, the complexification of utterances, and the emergence of features of discourse and cohesion, for instance, required considerably more attention and that register was a dimension that was of both developmental and practical importance and had to be included. Fifth, the FSI notion of "absolute" language proficiency levels in which the descriptors are checklists was considered unsatisfactory for reasons already stated in Section I above.

3 Selected issues

The nature and use of the ASLPR has been extensively discussed elsewhere [e.g., Ingram 1985, 1982, 1982b, 1980, 1979] but here a few key issues have been selected for consideration, some of which give the ASLPR its distinctive characteristics.

3.1 Parameters of Change

Each descriptor seeks to provide a fairly comprehensive picture of a learner's language at that level in each macroskill but different parameters of change are dominant at different levels. This is a result, not only of the general path of language development but also of the fact that language is learned in response to need. At the lowest proficiency levels, 0 to I-, the learner's language is largely formulaic consisting of rote memorized utterances learned in response to need, whether that need is the need to survive in the second language community or the need to cope with the situations created for the learner by the teacher. The dominant parameter of change is the range of situations that the learner can cope with but, as the learner's proficiency approaches 1, some limited creativity starts to emerge as the learner uses whatever resources he or she has in order to produce original utterances to cope with new situations. Once

immediate survival needs are met, the learner will start to reach out to other people, to establish social relations, and, hence, to express personal as against universal meaning. Thus, from about 1+ to 2+, the dominant parameter of change is the complexification of the language with the emergence of those complexifying features (embedded phrases and clauses, and the features of cohesion and discourse structure) that help to make meaning precise. It is significant that this process, the linguistic forms that occur, and the motivation for development all have their parallels in the depidginization process that Schumann, Bickerton and others describe [e.g., Schumann 19791. As the learners' language base increases, so they can choose more deliberately and discriminatingly from within their language repertoire to modify their language to meet situational requirements and to extrapolate meaning for new items and so, from 3 to 4, the dominant parameter of change is register sensitivity and flexibility. Beyond this, from 4 to 5, the dominant parameter of change is the ability to see beyond surface meaning, to comprehend and use linguistically and culturally based humour and irony, and to draw on significant cultural understanding.

3.2 General Proficiency

The ASLPR in its present form is designed to measure general proficiency. This raises a number of significant issues. First, since language occurs only in situations and the situations in which it occurs determine the language forms that occur, it could be argued that one cannot speak of "general proficiency" so much as proficiency in this situation or that, in this register or that, and that one can speak only of specific purpose proficiency. However, such a view seems, at best, to be counter-intuitive since, if we say that X speaks Chinese and Y speaks Motu, we do not mean that X can only give a lecture on neurosurgery or that Y can only describe the parts of an outboard motor. Rather, when we say that someone can speak a language we mean that that person can speak the language in the sorts of everyday situations that people commonly encounter, the sorts of situations that human beings living in a physical and social world must necessarily encounter. Hence general proficiency refers to the ability to use the language in these everyday situations. In addition, even though there are observable differences between registers (in, for example, lexis, frequency of certain syntactic forms, discourse structuring, phonology, and so on), no register is entirely discrete or it would not be recognized as part of the language and the concept of general proficiency entails the ability to use those commonly occurring features of the language that emerge frequently across registers.

Second, one has to distinguish the underlying general proficiency from its realisation in a particular task in a particular situation. Thus, for instance, someone who has never used public transport nor attempted to read a bus timetable, may have difficulty doing so even though his or her reading proficiency might be quite high (e.g. 3). The ASLPR seeks to measure the underlying general proficiency rather than the learner's ability to

use the language in this particular task or in that particular situation. For this reason, the ASLPR is not accompanied by a set of fixed test materials but interviewers are encouraged to select their own interview material using the scale and especially the Example column to guide them in what materials are appropriate to particular levels and they are encouraged to vary the activities in order to establish whether the learner fails to carry out a task merely because it is novel or because his or her underlying proficiency is inadequate. In practice, this means that interviewers must develop and skillfully use their own set of interview materials and that there is a considerable emphasis placed on the training of interviewers and raters rather on the normal validation and standardisation processes associated with test development.

3.3 Proficiency not Communicative Competence

The ASLPR claims to measure proficiency and not communicative competence in the loose sense of the ability to communicate. The latter depends on things other than language and the ability to use it, including intelligence, education, general knowledge, introversion or extroversion, and the willingness of the other interlocutor to accommodate the learner's less than native-like utterances. The ASLPR can be used to rate language proficiency in the sense of the ability to mobilize language forms to carry out communication tasks. It includes notions of the sorts of tasks that can be carried out and how they are carried out.

4 Uses of the ASLPR

The ASLPR provides a description of how language proficiency develops from zero to native-like. Therefore it provides an overarching framework within which language programmes may be coherently planned with their entry levels and goals specified in proficiency terms. In addition, it can be used to state a learner's proficiency, i.e., to state the sorts of tasks a learner can carry out and how they are carried out. Hence, the ASLPR can be used to measure a learner's language proficiency, to assess the level of proficiency required to carry out particular tasks, and so, for example, to identify the level of proficiency for a particular vocation or it can be used to classify, for example, library materials or teaching resources. In brief, some of the uses (amongst many others) to which the ASLPR has been put include:

* as an integrating framework within which an integrated series of ESL courses are planned in English language teaching centres in Australia, (e.g., the Institute of Applied Linguistics in the Queensland University of Technology and the Centre for Applied Linguistics and Languages in Griffith University in Brisbane
* to state entry proficiency levels, exit goals, and attainment levels in courses;

* to stream learners into classes (though proficiency should be seen as only one of the issues to be considered for streaming);
* to assess the adequacy of applicants' language skills for vocational registration purposes (e.g., the Board of Teacher Registration in Queensland specifies in ASLPR terms minimum English language proficiency levels for overseas trained applicants for teacher registration);
* to assess the legal responsibility of defendants in a court of law who had signed police-prepared confessions [e.g., Ingram 1980]
* to assess the foreign language proficiency attainments of Secondary School students in competency-based approaches to syllabus design;
* to specify the minimum proficiency levels for foreign language teachers in Queensland schools [e.g., Ingram and John 1990]
* to develop data on rate of change of proficiency in language classes as a basis for more rational planning and resource allocation.

5 Assessing with the ASLPR

Assessment with the ASLPR, as with most proficiency rating instruments, is a matter of comparing learners' observed language behaviour with the scale descriptions and rating the learners' proficiencies at that level whose description best matches the observed language behaviour. Characteristically, the learner's maximum language behaviour is elicited in an interview though other activities may (as with the International English Language Testing System or IELTS) be used for the purpose. The ASLPR is deliberately not accompanied by a set of "standardized" or even recommended test activities for the reasons already indicated, i.e., because language is significantly situation-dependent, it is necessary to elicit language behaviour in situations that are relevant to the learner or at least to ensure that the learner's inability to perform in a situation is not the result of unfamiliarity with that situation rather than the level of language development he or she has attained. The ASLPR provides some guidance as to tasks that are relevant to learners at different levels through the "Example" column and training programmes provide training to interviewers in the selection and preparation of interview materials.

The aim of the interview is to elicit the learner's maximum language behaviour. Consequently, the interview is in three broad phases. In the first, the exploratory phase, the interviewer is required to settle the learners down, gain an approximate idea of where on the scale their proficiency is likely to fall, and to decide what topics it may be productive to pursue. By the end of this phase, the interviewer should have an idea of the approximate two or three proficiency levels around which their proficiency in the macroskills in question is likely to fall. The next phase, the analytic phase, is the one in which the interviewer deliberately guides the interview in order to explore the learners' language, leads them to their maximum language behaviour (i.e., their language ceiling or linguistic breaking-point), and clarifies which of the proficiency levels best

describes their language behaviour. In the final or concluding phase, the interviewer is advised to drop back to a level within the learners' proficiency range in order not to send them away with a feeling of failure having been led to their "linguistic breaking-point".

Behaviour in all four macroskills is elicited in essentially the same way and so, in the literature on the ASLPR, one generally speaks only of interviewing because the same approach applies to all four macroskills. The manner in which one observes the learners' language behaviours is, however, a little different in each case. With Speaking, the rater is essentially observing the learners' overt language behaviour. In Reading and Listening, receptive skills, that behaviour is internal and one uses answers to questions and other stimuli to deduce that behaviour. Writing is somewhat similar to Speaking except that it is more economical of time for the writing test to be done in a group and for the rater to use the learners' answer scripts to deduce the writing behaviour that has gone on.

6 Trialling

6.1 Overview

In assessing the validity and reliability of direct instruments, there are at least three issues to be considered: the adequacy of the scale itself (i.e., the extent to which it can be regarded as adequately representing both the overall path of language proficiency development and a series of coherent levels along that path), the extent to which interviews of the prescribed type can regularly elicit the learners' maximum language behaviour for rating purposes, and the extent to which raters can validly and reliably rate the learners' proficiencies.

In trialling the ASLPR, the second issue received only limited attention partly because there has been much research into the interview process and the ASLPR's interview process is little different from that of the FSI Scale and also because any study of this issue must be dependent on the validity and reliability of the scale itself, which, therefore, had to be examined first. The first and third issues tend also to merge since it was assumed in the trials that, if raters were to be able to rate validly and reliably, the scale had to provide a valid description of how language develops from zero to native-like and each descriptor had to be coherent and provide a valid description of a real stage of development. If the descriptors were not so, then one would expect that raters would be inconsistent in their rating of the learners because they would be focusing on different features out of the jumbled descriptors. Consequently, the main aim of the formal trials of the ASLPR was to assess the extent to which different raters were able to interpret and apply the ASLPR similarly to different learners and, in particular, to interpret and apply the scale in the same way as do its authors.

Consequently, some sixteen learners (two at each Speaking level from 0+ to 5) were interviewed onto video demonstrating their Speaking,

Listening, and Reading proficiencies and Writing scripts were collected for them (each learner having been asked to do the same writing exercises that led them from a simple transcription exercise through simple notes to quite complex pieces of correspondence). The learners' proficiencies were then rated by groups of trained ESL teachers in Australia (native English speakers) and China (non-native speakers of English) and their ratings compared with each other and with those assigned by the authors of the ASLPR. The Australian teachers were asked to rate the same videos and written scripts on two different occasions approximately 12 months apart. In addition, the proficiencies assigned to the learners using the ASLPR were compared with results they obtained on the CELT test [Harris and Palmer 19701, some close passages and a dictation. Subsequently approximately 200 Second and Third Year EFL students in a foreign language institute in China were also rated on the ASLPR and asked to complete the CELT, cloze and dictation tests. The results of these trials are written up in detail elsewhere [Ingram 1982] and are summarized here. Subsequently, the same videos and written scripts were rated by a group of American teachers (i.e., speakers of a different dialect of English from the target variety being learned by the learners on the videos). Similar trials were also held using adolescent learners of English as a Second Language, adult learners of English as a Foreign Language, and adolescent and adult learners of French, Italian and Japanese as Foreign Languages.

6.2 Validity

Face validity seems to be indicated by the widespread acceptance of the ASLPR (especially in Australia but also elsewhere) and the extent to which it is now used both in assessing learners' proficiencies and for the other purposes outlined earlier. Construct and content validities were continuously assessed during the development process as scores of learners of English and other languages were interviewed and observation of their language used to review, elaborate and refine the descriptors of each proficiency level. In addition, as suggested earlier, in assessing inter- and intra-rater reliabilities, one is assessing at least content validity since, if the descriptors of each level were not coherent or failed to reflect actual points in the common developmental schedule, one would expect that raters would differ in their interpretation of the scale and in the elements focused on in different learners' performances, and that the ratings they assigned would not be consistent. This was not the case and, as will be seen later, there was a high degree of inter- and intra-rater reliability.

Concurrent or immediate pragmatic validity was assessed by comparing scores on the ASLPR with scores on CELT, clozes and dictations. Although concurrent validity is commonly included when one is assessing a new test or course, it involves significant logical problems; on the other hand, the results may provide interesting insights into the usefulness of tests in different learning contexts and, in the case of a behavioural rating scale compared with indirect and semi-direct tests, if the level of concurrent validity should be uniformly high and the ratings assigned re-

liable, the rating scale could be used to explicate the scores received on the other tests.

The statistical analysis of the data in the ASLPR trials involved the use of a variety of routines in the Statistical Package for the Social Sciences run on a Hewlett Packard HP-3000 Series III processor. Routines used were condescriptive, Pearson correlation, and non-parametric (Spearman Rank Order) correlation [Nie et al 1975: 181, 216, 2881. Additional routines included canonical correlation and scattergrams based on the teachers' mean ratings. The arguments for and against the use of these procedures and their applicability to the type of tests used in this study are discussed in the detailed report of the trials [Ingram 1982].

The ASLPR ratings assigned by the developers of the scale were compared with results on the CELT Test [Harris and Palmer 1970], the latter being chosen because of its ease of availability and the fact that Carroll reports a high degree of similarity between its results and those of TOEFL [reported in Buros 1975: 207]. ASLPR ratings were also compared with results on three clozes and three dictations (easy, medium and hard in each case).

Table 1 shows the correlation coefficients between the ASLPR ratings and CELT sub-tests for ESL learners in Australia. These range from r = 0.8 (ASLPR Speaking and CELT Structure) to rho = 0.96 (ASLPR Reading and CELT Total) with significance levels at or beyond the .001 level. If anything, these coefficients are slightly higher than in similar studies of the FSI Scale and other standardized tests reported in the literature to that time [e.g., Mullen 1978, Clifford 1978, Carroll 1967, Politzer et al 19821. However, for foreign language learners in China, the correlation coefficients were very different. For one group (Table 2), they ranged from rho = 0.01 (ASLPR Reading and CELT Vocabulary) to rho = 0.59 (ASLPR Speaking and CELT Structure) with generally moderate significance levels while, for the other group (Table 3), they ranged from rho = -0.25 (ASLPR Reading and CELT Listening) to rho = 0.62 (ASLPR Writing and CELT Listening) with significance levels up to 0.003 but generally very low.

These ambivalent results showing different levels of correlation for different groups of learners are similar to those found in other studies of direct and indirect tests [e.g., Mullen 1978] and seem not to arise so much from any instability inherent in the ASLPR itself as from factors relevant to the way in which the language has been learned and the relationship between analytic, especially knowledge-based, tests and practical proficiency. Learners in a second language situation have continual experience of applying their knowledge in everyday communication activities and so one might expect that their active communication abilities and their passive knowledge of the language would be more alike than for learners in a foreign language situation (especially where traditional grammar-based courses are being followed). In a foreign language situation, one would expect more difference between formal knowledge and practical ability since the learners have fewer opportunities to practise their language in real communication activities and the extent of the relationship would seem likely to depend on the extent to which the individual learner was

willing to "have a go", his ability to make use of whatever language resources he or she has available, and the extent to which he or she has taken advantage of whatever language use opportunities there might be both inside and out of the classroom. This was further borne out during the later trials of the French, Italian and Japanese versions of the ASLPR: here all final year Secondary School students had basically covered the same language content but those who had the opportunity to use a variety of languages in their everyday life demonstrated higher levels of practical proficiency.

Table 1

Correlations between ASLPR Macroskill Ratings and the Comprehensive English Language Test (CELT) for Adult Learners of ESL in Australia

ASLPR	CELT								
	Structure-A		Listening-A		Vocabulary-A		CELT	Total	
n	19		18		19		18		
	r	rho	r	rho	r	rho	r	rho	
Speaking	.83	.91	.86	.90	.88	.87	.89	.90	
p	.000	.001	.000	.001	.000	.001	.000	.001	
Listening	.80	.90	.86	.90	.85	.83	.87	.88	
p	.000	.001	.000	.001	.000	.001	.000	.001	
Writing		.88	.94	.89	.90	.91	.91	.93	.95
p	.000	.001	.000	.001	.000	.001	.000	.001	
Reading	.84	.95	.87	.90	.92	.94	.92	.96	
p	.000	.001	.000	.001	.000	.001	.000	.001	

rho = Spearman Rank Order Correlation Coefficient
r = Pearson Product-Moment Correlation Coefficient

Table 2

Correlations between ASLPR Macroskill Ratings and the Comprehensive English Language Test (CELT) for Third Year Students of EFL at the Guangzhou Institute of Foreign Languages, Guangzhou, China

ASLPR	CELT								
	Structure-A		Listening-A		Vocabulary-A		CELT Total		
n	21		21		21		21		
	r	rho	r	rho	r	rho	r	rho	
Speaking	.57	.59	.50	.52	.30	.26	.51	.55	
p	.003	.002	.01	.008	.094	.13	.009	.005	
Listening	.52	.55	.50	.52	.32	.27	.51	.58	
p	.008	.005	.01	.008	.081	.121	.01	.003	
Writing		.44	.49	.64	.57	.51	.038	.64	.57
p	.022	.012	.001	.003	.01	.043	.001	.003	
Reading	.32	.43	.41	.45	.14	.01	.31	.30	
p	.079	.025	.031	.021	.274	.483	.083	.096	

Table 3

Correlations between ASLPR Macroskill Ratings and the Comprehensive English Language Test (CELT) for Second Year Students of EFL at the Guangzhou Institute of Foreign Languages, Guangzhou, China

ASLPR	CELT							
	Structure-A		Listening-A		Vocabulary-A		CELT Total	
n	20		20		20		20	
	r	rho	r	rho	r	rho	r	rho
Speaking	.10	.22	-.06	.02	.10	.18	.04	.19
p	.33	.18	.40	.46	.50	.23	.44	.21
Listening	.38	.47	.16	.20	.40	.47	.39	.50
p	.05	.02	.25	.20	.04	.02	.04	.01
Writing	.62	.60	.61	.62	.30	.30	.58	.58
p	.003	.002	.003	.003	.114	.114	.006	.005
Reading	.08	.10	-.19	-.25	.13	.07	.03	.002
p	.38	.3	.21	.14	.29	.39	.46	.50

6.3 Reliability

In trialling a proficiency rating scale, reliability is, for reasons already implied, probably of most interest. Two questions were principally asked:

1. To what extent can teachers, whether native or non-native speakers of English interpret and apply the scale in the same way as do its authors (inter-rater reliability)?
2. To what extent can teachers interpret and apply the scale in the same way, to rate the same learners, on two separate occasions (intra-rater reliability)?

These questions were important for two reasons, first, because, if a scale is to be used by many different people, one wants to be sure that they can all interpret it similarly and, second, because, as already indicated, if the scale failed to provide a true and coherent picture of how proficiency develops and of identifiable levels within the schedule of proficiency development, then raters would tend to focus on different features more or less haphazardly with regard to the descriptors and so one would expect a low level of relability.

Reliability was assessed by the procedure outlined earlier (Section VI.1). Tables 4 to 6 show the correlation coefficients when each teacher's ratings for the 16 learners are compared with the authors' ratings. The correlations for inter-rater reliability are generally in excess of 0.9 with means from 0.94 to 0.96 for the Australian teachers and with means of 0.93 to 0.95 for the Chinese teachers (who also exhibited a slightly greater range). Intra-rater reliabilities for the Australian teachers were also generally in excess of 0.9 with means of 0.96 or 0.97. Clearly the correlations are high and suggest that all the teachers interpreted and applied the scale in essentially the same way as did the scale's authors and, in the case of the Australian teachers, they did so on two different occasions a year apart. Though the levels of correlation with the authors' ratings were slightly lower for the non-native English speaking Chinese teachers, they are still quite high and suggest that even non-native speakers can use the scale reliably.

Table 4

Inter-rater Reliability: Australian Teachers and the Authors of the ASLPR

Macroskill	rho			r		
	min.	max.	mean	min.	max.	mean
Speaking	.91	.99	.96	.94	.99	.96
Listening	.89	.98	.94	.90	.99	.94
Writing	.93	.997	.96	.93	.99	.96
Reading	.91	.99	.96	.91	.98	.95

p< .001 for all rho's; p .001 for all r's

Table 5

Macroskill	rho			r		
	min.	max.	mean	min.	max.	mean
Speaking	.88	.98	.95	.83	.98	.93
Listening	.88	.96	.93	.89	.97	.94
Writing	.88	.98	.95	.87	.98	.94
Reading	.88	.98	.95	.83	.98	.93

$p < .001$ for all rho's; $p \leq .001$ for all r's

Table 6

Macroskill	rho			r		
	min.	max.	mean	min.	max.	mean
Speaking	.93	.99	.97	.93	.99	.97
Listening	.91	.98	.96	.92	.98	.96
Writing		.92	.98 .96		.91	.99 .96
Reading	.87	.98	.96	.91	.98	.96

$p < .001$ for all rho's; $p \leq .001$ for all r's

In brief, although only a brief analysis of the results of the formal trials of the ASLPR is provided here and interested readers are recommended to consider the more detailed results and analyses in Ingram 1982, these results do suggest that acceptable levels of validity can be claimed and high levels of inter- and intra-rater reliability. Thus, the ASLPR does seem able to make valid statements about learners' language proficiency and users of the ASLPR, whether native or non-native English speakers, do seem able to interpret and apply the scale reliably and to make valid and reliable assessments of learners' language proficiencies. The results also suggest that the often-heard criticism of "direct tests" that they are unduly subjective and, therefore, necessarily unreliable is ill-founded. It is important, however, to emphasize that all persons participating in the trials of the ASLPR had been provided with a training programme in the use of the ASLPR; indeed, the authors of the ASLPR always, as a matter of principle in the use of direct proficiency assessment instruments, emphasize the importance of the training of both interviewers and raters.

References

Australian Department of Immigration and Ethnic Affairs (1984).
Australian Second Language Proficiency Ratings. Canberra: Australian Government Publishing Service.

Buros, Oscar K. (ed) (1975).
Foreign Language Tests and Reviews. Highland Park, NJ: Gryphon Press.

Carroll, John B. (1967).
The Foreign Language Attainments of Language Majors in the Senior Year: A Survey conducted in U.S. Colleges and Universities. Cambridge, MA: Harvard University Graduate School of Education.

Clark, John L. D. (ed) (1978).
Direct Testing of Speaking Proficiency: Theory and Application. Princeton, NJ: Educational Testing Service.

Clark, John L. D. (1972).
Foreign Language Testing: Theory and Practice. Philadelphia, PA: Center for Curriculum Development.

Clifford, R. J. (1978).
Reliability and Validity of Language Aspects Contributing to Oral Proficiency of Prospective Teachers of German. In: Clark, 1978, 191-209.

Foreign Service Institute School of Language Studies (1968).
Absolute Language Proficiency Ratings. mimeograph. Reprinted in Clark 1972:122-123.

Harris, David P. and Leslie A. Palmer (1970).
A Comprehensive English Language Test for Speakers of English as a Second Language. New York: McGraw-Hill.

Hyltenstam, Kenneth and Manfred Pienemann (1985).
Modelling and Assessing Second Language Acquisition. Clevedon: Multilingual Matters.

Ingram, D.E. (1978).
An Applied Linguistic Study of Advanced Language Learning. Unpublished thesis for the degree of Doctor of Philosophy, University of Essex, Colchester, 1978. In ERIC Collection ED168 359.

Ingram, D.E. (1979).
An Introduction to the Australian Second Language Proficiency Ratings. Mimeograph. Also in Australian Department of Immigration and Ethnic Affairs 1984: 1-29.

Ingram, D.E. (1980).
Proof of language incompetence. In: *The Linguistic Reporter*, 23, 1, 14-15.

Ingram, D.E. (1982a). English language proficiency in China: A study.
MLTAQ Journal, 17, 21-45.

Ingram, D.E. (1982b).
New Approaches to Language Testing. Paper to the Fourth National Language Conference of the Australian Federation of Modern Language Teachers Associations, Perth, 3rd-6th September, 1982. Reprinted in MacPherson 1982.

Ingram, D.E. (1983).
Report on the Formal Trialling of the Australian Second Language Proficiency Ratings (ASLPR). Canberra: Australian Government Publishing Service, 1983.

Ingram, D.E. (1985). Assessing Proficiency: An Overview on Some Aspects of Testing. In: Hyltenstam and Pienemann, 1985: 215-276.

Ingram, D. E. and Glyn John. (1990).
The Teaching of Languages and Cultures in Queensland: Towards a Language Education Policy for Queensland Schools. Brisbane: Griffith University Reprographics.

Ingram, D.E. and Elaine Wylie. (1979, revised 1982, 1985).
Australian Second Language Proficiency Ratings. Mimeograph. Also in Australian Department of Immigration and Ethnic Affairs 1984: 31-55. (First edition of ASLPR published 1979, current edition November 1985).

Ingram, D.E. and Elaine Wylie. (1982).
Australian Second Language Proficiency Ratings (Version for French as a Foreign Language). Brisbane: Brisbane College of Advanced Education, Mount Gravatt Campus*.

Ingram, D.E. and Elaine Wylie. (1982a).
Australian Second Language Proficiency Ratings (Version for Italian as a Foreign Language). Brisbane: Brisbane College of Advanced Education, Mount Gravatt Campus*.

Ingram, D.E. and Elaine Wylie. (1982b).
Australian Second Language Proficiency Ratings (Version for Japanese as a Foreign Language). Brisbane: Brisbane College of Advanced Education, Mount Gravatt Carnpus*.

MacPherson, Anne (ed) (1982).
The Language Curriculum in the 1980's Affective, Effective or Defective? Perth: MLTAWA / AFMLTA.

Mullen, K. A. (1978).
Determining the Effect of Uncontrolled Sources of Error in a Direct Test of Oral Proficiency and the Capability of the Procedure to Detect Improvement following Classroom Instruction. In: Clark, 1978:171-189.

Nie, Norman H. et al. (1975).
Statistical Package for the Social Sciences. New York: McGraw-Hill.

Politzer, Robert L., Elana Shohamy, and Mary McGroarty (1982).
Validation of Linguistic and Communicative Oral Language Tests for Spanish English Bilingual Programs. Paper to the Pre-Conference Symposium on Language Testing, 1982 TESOL Convention, University of Hawaii, Honolulu, Hawaii. Mimeograph.

Schumann, John H. (1979).
The Pidginization Process: A Model for Second Language Acquisition. Rowley, Mass.: Newbury House

*Now Mount Gravatt Campus, Griffith University.

CROSS-NATIONAL STANDARDS: A DUTCH-SWEDISH COLLABORATIVE EFFORT IN NATIONAL STANDARDIZED TESTING

John H. A. L. de Jong,
CITO, Dutch National Institute for Educational Measurement

Mats Oscarson,
Gothenburg University

1 Introduction

In most comparability studies different tests are compared by administering them to samples from the same population. One of the major new insights from modern test theory is that a comparison can start out from either the items in a test or the persons in a sample. By comparing the results of two nominally distinct populations on the same test, relevant information can be gathered about the appropriateness of the test as a population-independent measurement instrument suitable for transnational or international assessment designs.

The study reported here was conducted to investigate the possibility to use data gathered on tests of reading comprehension of English as a foreign language from a sample of Swedish students, to predict item and test characteristics if the tests were to be used in the Netherlands. This objective was primarily pragmatic. The Dutch educational authorities object to the pretesting of examination papers for fear of undue disclosure. If it was possible to pretest examinations abroad, the risk of disclosure would be reduced to a negligible level. Secondly, we were interested from an applied-linguistics point of view, to investigate the scalability of a foreign language skill across different language backgrounds and given different curricula.

The Swedish and Dutch educational systems differ in many aspects, which could lead to quite different knowledge and skill domains within the two populations of students. However, it has been suggested that at higher levels of language proficiency, subjects become independent of the curriculum and their proficiency can be assessed on a single dimension (De Jong, 1987; De Jong & Henning, 1990) Before even attempting to answer the main research question of this study, it is therefore necessary to investigate the effectiveness of tests designed for examination use in the Netherlands as measurement instruments to distinguish levels of English proficiency among Swedish students.

The study was to address questions pertaining to psychometric indicators of test and item quality. For pretesting purposes data are needed that allow predictions to be made at several levels. At the level of tests it is less difficult to obtain acceptable precision than at the level of individual items. The questions we set out to answer are given below in an order of increasing predictive power necessary.

- Can a score transformation function be defined such, that average (sub)test scores as observed in Sweden can be used to predict average (sub)test scores to be obtained by students taking their school leaving examinations in the Netherlands?
- Can differences in difficulties between items estimated from test administration in Sweden be used to predict differences in difficulties between items if estimated from administration in the Netherlands?
- Can differences in item quality as apparent from item-rest correlations observed in Sweden be used to predict differences in quality if the items were taken by Dutch students?
- Can differences in distractor attractiveness as observed from test administration in Sweden be used to predict distractor attractiveness for Dutch students?

2 The Swedish and Dutch educational systems

To enable the reader to judge on the relevance of our study, we provide a brief description of the Swedish and Dutch educational systems, in particular with respect to the teaching and testing of English as a foreign language in upper secondary education.

2.1 Sweden

Compulsory education in Sweden starts at the age of seven, when children after two optional years of preschooling enter the Grundskolan (basic school'). The Grundskolan is a comprehensive type of school and has a nine year program. It covers primary education and lower secondary education. Two or three further years of general education are offered in the Gymnasieskolan (upper secondary school). About 35% of the age cohort opts for one of the following lines of study (streams) in the three-year course: Humanistisk linje (humanities), Naturvetenskaplig linje (natural science), Samhdllsvetenskaplig linje (social sciences), Ekonomisk linje (economics), or Teknisk linje (engineering). These lines of study represent the regular option for students aiming at higher education at the university or college level.

English is studied as the first foreign language by all Swedish pupils. It is introduced in grade 3 or 4 (age 9 or 10) and remains compulsory until the end of Grundskolan (end of grade 9, age 16). At that time students will have had some 400 hours (net) of English. In the Gymnasieskolan most students have another 200 hours of English. During the last three years of Grundskolan students can opt for a second foreign lan-

guage, German or French. The second foreign language, referred to as 'language B', is taken by 60-65% of the students for a total of approximately 250 hours. In the Gymnasieskolan a third language (language Q can be chosen from German, French, Spanish, Italian, Russian, or Finnish.

To relieve the pressure of end-of-school testing, Sweden replaced external final examinations in the late 1960's by a system of continuous internal school assessment. However, in order to maintain a framework for equivalent marking and grading throughout the country, national standardized tests covering the most important subjects, such as mathematics, physics, chemistry, Swedish, English, French, and German, are administered at a number of stages during the curriculum. The tests conducted in the Gymnasieskolan, the so-called Centrala prov ('national tests') are compulsory. In languages they are scheduled in the spring of the second year (grade 11), after about 500 to 550 hours of English and 300350 hours of 'language B'. The primary aim of the national testing program is to provide information to the schools about their position in relation to the national average. The implicit philosophy is that teachers are perfectly capable of ranking rank their students within their own classes. But they have no information on the location and spread of their students' ability with respect to the national standards. The raw scores of random samples of a few thousand students are gathered nationally and standardized on a five point scale, yielding a mean of 3.0 and a standard deviation of 1.0. Thus in the first year of the upper secondary school, for instance, the grades I through 5 are awarded to 7, 24, 38, 24, and 7 per cent respectively of the total population taking the same course in a given year. Norm data and general information about test results are reported to the schools and are to be used as guidelines by the teachers in grading their students at the end of the term or the academic year. In this way the teacher marks in each class are adjusted to take into account the performance of the class on the national tests and grades given within the schools can be taken to have approximately the same meaning across the country. Finally, it is important to note that pass-fail decisions are largely unknown in Swedish schools. The system operates on the basis of self-selection (for more details see National Swedish Board of Education, n.d.).

Generally the foreign language tests in the Centrala prov have consisted of four subtests, viz., grammar, vocabulary, reading comprehension, and listening comprehension. For a number of years the tests also included an optional written-production part. In the future an oral part is likely to be added. The compulsory part of the test generally uses the multiple-choice and fill-in-the-blank formats (see also Lindblad, 1981; 1983; Orpet, 1985; Oscarson 1986).

2.2 The Netherlands

Compulsory education in The Netherlands starts at the age of four, when children enter the Basisschool ('basic school'). This is a primary school, which includes two years of pre-school and has an eight year program.

After finishing the Basisschool children can choose from four different streams of secondary education: 'pre-university education' (Dutch acronym: VWO, six years), 'higher general secondary education' (HAVO, five years), 'lower general secondary education' (MAVO, four years), and 'vocational education' (LBO, four years). A total of about 36% of the age cohort opts for VWO (16%) or HAVO (20%) which represent the regular options for students aiming at higher education at the university or college level respectively.

In 1986 English was introduced as the first foreign language for all Dutch pupils in grade 7 (age 10) of primary school. The Dutch students in this study, however, had finished primary school before 1986 and started English in the first grade of secondary school (age 12). They started French at the same time, and German one year later. The three languages remain compulsory through grade 4 of VWO and grade 3 of HAVO. From then on students choose a minimum of 7 school subjects in VWO and 6 subjects in HAVO. At least one foreign language is compulsory, many students take two foreign languages and a minority takes three languages. About 96% of the students takes English, 40% German, 25% French, and less than 5% chooses Spanish or Russian. At the time of their final examination students who have chosen e.g. English will have had about 600 hours of instruction in VWO schools and about 550 hours in HAVO schools.

Examinations at the conclusion of secondary education in the Netherlands consist of two parts: a national external examination and continuous school internal assessment. Each part contributes 50% to the final mark. The subject matter covered by the constituent parts varies. For some school subjects, such as mathematics and chemistry, both parts cover the same matter; the difference then lies in the mode of assessment. For the foreign languages the external part deals with reading comprehension only, whereas speaking, writing, listening, and literature are assessed by the schools. All parts of the external and internal assessment are graded between 1.0 (minimum) and 10.0 (maximum), where 5.5 is the minimum passing mark, On the examination papers in foreign languages students obtain about 6.5 on average with a standard deviation of 1.1. On most tests used in Dutch education raw scores are transformed to examination marks by a system which is a mixture of criterion-referencing and norm-referencing. Before test administration a group of experts defines a score-range for the cut-off. After test administration the cut-off is fixed to fall within this range, but in such a way as to keep the proportions of pass/fail decisions per subject more or less constant over subsequent cohorts. In the area of foreign language testing some experiments have been carried out with respect to equating the external examinations, but the observed distribution is still used in determining the cut-off. However, for CITO listening comprehension tests, a standard, cohort-independent equating procedure is operational since 1985. The percentage of fails on any given subject matter ranges from a low 5% (Dutch as L1) to a high 40% (mathematics). For the foreign languages the average number of fails is about 20% on any paper in both the external and internal examination.

Because the pass/fail decision on the total examination is comp over all subject matters and students are allowed one or two fails, the success rate is over 90%.

The official examination program for the foreign languages, laid down in the 1968 Education Act, is aimed at communicative language learning, it mentions the modes of language use and some knowledge of the literature. However, in actual practice, a large minority of teachers continues to assess grammar and vocabulary separately in the school internal examination. The external examination is a multiple-choice reading comprehension test. The internal examination differs from school to school, but mostly consists of about 3 or 4 written papers and 2 oral sessions testing for the other modes of language use. There is no moderation system for the internal assessment, but schools have to submit their assessment procedures for evaluation to national inspectors of education. Furthermore, for some components national standards do exist because many teachers use tests made available by the National Institute for Educational Measurement (CITO). For instance, almost all teachers use the CITO tests to assess their students' foreign language listening comprehension.

3 Method

3.1 Subjects

The Swedish pilot groups

A total of 361 students (16 classes) took part in the administration of the Dutch reading tests in Sweden. The Swedish students were all in their third year of the upper secondary school (Gymnasieskolan) and were from six schools in the mid-western region of the country. The average age of the subjects was 17.5. They represented four of the lines of study mentioned in 2.1 above (humanities, natural science, social science, and engineering), and hoped to continue further study at the university and college level in the next academic year. The subjects were all in their ninth year of English as a foreign language and had had a total of some 500-600 hours of instruction (net) when they took the tests.

The Dutch students

Reading comprehension tests are administered as part of the examination procedure. On a regular basis data are gathered on all students taking the tests (some 34,000 students in the VWO-stream and 41,000 in the HAVOstream). After computing the main test statistics, further analyses are carried out on a random sample of little over 1,000 out of each total population. The average age of the subjects is 18 for VWO-students and 17 for HAVO-students. At the time of the examinations VWO-students are in their sixth year of English and have had ±600 hours of instruction, HAVO-students are in their fifth year and have had ±550 hours of English.

3.2 Instruments

Each year three new test forms of each examination paper are constructed following a standardized procedure. There is one form for the regular session of the examination and two for resits. The National Examination Board (Dutch acronym: CVO) is responsible for their production, which is carried out under strict security regulations. Texts for use in the examinations are sent in by teachers and evaluated by CITO staff. After approval of the texts, test items are conceived by teachers and are submitted to several rounds of criticism. Written comments are evaluated and discussed in meetings chaired by CITO staff. The final version has to be approved by the National Examination Board before it is sent to the national printer.

The foreign reading comprehension examinations for VWO and HAVO consist of five texts of ±700 words each, which are original essay-type articles reproduced from newspapers, magazines, etc., published in the target language for a native speaker readership. Each text is accompanied by ±10 multiple-choice questions, the total test length is 50 items. The items test for comprehension of words and structures in context, relations, literal and implied meaning, author's intent, sampled from an agreed matrix of reading skills and language functions. The texts are meant to increase in difficulty from text I to text 3 and to decrease in difficulty from the third to the last text. There is no attempt at ordering items within a text.

For this study we used the tests for the first session of the 1988 final external examinations of English reading comprehension for VWO and HAVO, comprising a total of 100 items (50 each) distributed over 10 subtests.

For a large proportion of the Swedish students in our study we were also able to collect the grades they obtained for English on an-end-of-term school internal test and on the Centralt prov for English, which would allow a comparison between the regular assessment of students' English proficiency in Sweden and their assessment using the Dutch examinations.

3.3 Procedures

3.3.1 Test administration

In the Netherlands the tests had been administered as the regular examinations of English reading comprehension to all Dutch students finishing HAVO and VWO schools. Random samples of the 1988 population of examinees, 1,014 from VWO and 1,016 from HAVO were drawn for the test analyses.

For the administration in Sweden three test booklets were compiled. Booklet 1 contained a single subtest taken from the 1988 HAVO examination and was meant to familiarize the Swedish students, if necessary, to the format of the tests. About 25-30 minutes are needed to answer the questions in a single subtest, thus leaving ample time for introduction and discussion within the same class period. The test was scored by the

students' own teachers, but the data were not used in this study. Booklet 2 and 3 each required a full class period as they contained two subtests, ordered according to difficulty. Booklet 2 started with subtest number 4 taken from the 1988 HAVO-exam (H-4, 10 items), followed by subtest number 2 from the 1988 VWO-exam (V-2, 11 items). Booklet 3 contained subtest number 5 from the 1988 HAVO-exam (H-5, 11 items), followed by subtest number 3 from the 1988 VWO-exam (V-3, 11 items). These subtests were used for the main part of the study.

As we planned to estimate item and subtest difficulty using modern test theory (Item Response Theory, IRT) it was not necessary that all students in the design took all subtests. The Swedish teachers participated in the study on a voluntary basis and spent as much time on the tests as they thought was appropriate. Therefore some students made only one subtest, some made two, some three. Three classes, with a total of 50 students, completed all four subtests. Table I provides an overview of the number of Swedish students from the four different streams that participated in the study for each subtest. Note that because students participated on average in 2.5 subtests, the numbers in the rows cannot be added to yield the total. Finally, students with more than one omit in their response record were removed from the analyses.

Table 1
Number of Swedish Students per Stream and per Subtest

	Per Subtest				Total in	Total in
Stream	H-4	H-5	V-2	V-3	design	study*
Humanities	61	48	61	11	66	57
Social Sciences	76	104	76	27	134	119
Natural Sciences	131	96	101	41	133	131
Engineering	0	28	0	28	28	26
Total in design	268	276	238	107	361	
Total in study*	247	256	217	99		333

* After editing out records with more than one omit

3.3.2 Data analyses

Test results were analyzed according to Classical Test Theory and to Item Response Theory. For the latter the one parameter Rasch model (Rasch, 1960) was chosen. Item parameters were estimated using a conditional maximum likelihood procedure. We used the computer program OPLM (Verhelst, Glas, & Verstralen, forthcoming) developed at the research department of CITO. OPLM is written for personal computers and performs CPU and memory efficient analyses of planned missing data designs, i.e., designs in which a 'test' is defined as a set of items tried by a particular group of subjects in such a way that each item in the 'test' is tried by all subjects in the group. Overlap exists between tests through items that fig-

ure in more than one test. OPLM can deal with dichotomous and polytomous items in any combination. Item parameters can be estimated using the conditional maximum likelihood (CML) or marginal maximum likelihood (MML) method. Given the estimate of the item parameters, the ability parameters of all groups were estimated using an unrestricted maximum likelihood method and a bias correction procedure suggested by Warm (1985) and implemented in OPLM by Verhelst and Kamphuis (1990). OPLM provides four goodness of fit tests for individual items (Verhelst & Eggen, 1989; 1990; Verhelst & Verstralen, forthcoming) as well as goodness of fit tests at the level of tests and of the total design (Glas, 1988; Glas 1989; for an application see De Jong & Glas, 1987).

Separate IRT analyses were run on the two target samples of Dutch students, the Swedish students and on all subjects from both nationalities combined. An additional analysis was run on the data from the Swedish students taking their grade on the Centralt prov as a single five-category item. This allowed the estimation of the relation between the variable measured by the 'English examinations' in both countries.

The quality of a multiple-choice item is influenced not only by the correct response, but also by the quality of the distractors. Distractors may be too attractive or may lack attractiveness. In our study, aimed at the investigation of possibilities for pretesting, it was important to assess the items on this particular aspect. Because of ceiling and floor effects it is less appropriate to compare observed proportions of examinees from two different populations that are attracted by each of the distractors, especially if the populations differ in ability. The same problem exists for the correct responses, but there it is solved by using IRT, which allows an estimation of item parameters independent of ability. For the estimation of distractor attractiveness no satisfactory procedures are available.

Nevertheless, the following argumentation was used to allow predictions concerning distractor attractiveness. In the Rasch model the total score is a sufficient statistic to estimate ability. When using a zero-one scoring rule for multiple-choice items, one implicitly assumes that counting the total number of correct responses given by a person is the best procedure to arrive at a total score for this person and that, given an incorrect response, no further useful information on ability can be abstracted from the person's choices between the distractors. If, on the other hand, one assumes that choosing a particular distractor is dependent on ability, zero-one scoring is not the most appropriate procedure and a different score should be awarded to each distractor. This would lead to the Partial Credit Model derived by Masters (1982), where item responses are scored in ordered categories. Therefore, the assumption that zero-one scoring is the best procedure, implies the assumption that, given an incorrect response, the choice of a particular distractor over the other(s) is independent of ability. Distractor popularity, therefore, must be a characteristic of the distractor in relation to the other distractor(s). Furthermore, the sum of the probabilities of selecting any one of the distractors is constrained by the probability of scoring zero. This leads to the simple conclusion that for each item the ratio of the probability of adopting a dis-

tractor and the probability of scoring zero is a constant, independent of ability. This fixed ratio then, can be used to estimate distractor popularity at any ability level from the observed proportion of subjects at any other ability level.

4 Results and Discussion

4.1 General results

Table 2 presents the statistics of the samples of Dutch students from the examination populations and the results of the Swedish students collected in this study. To facilitate comparisons between subtests differing in length, the table also provides an estimate of the reliability of all subtests at a common length of 50 items (KR-20*) using the Spearman-Brown formula. Apparently the 1988 HAVO examination had been well targeted on the intended population, the mean percentage score (Xperc) was halfway between the maximum score (100%) and the chance score (25%).

Table 2
Test Statistics for Total HAVO and VWO English Reading Comprehension Examinations (Dutch Students), and for Four Subtests (Dutch and Swedish Students).

Test	k	Sample	N	X(abs)	X(perc)		StDev	Se	KR-20 (*)
HAVO	50	Dutch 1014	31.35	62.70	7.48	3.04	.83	–	
H-4	11	Dutch 1014	6.66	60.55	2.19	1.46	.56	.85	
		Swedes	247	7.83	71.18	2.18	1.35	.61	.88
H-5	10	Dutch 1014	6.23	62-30	2.08	1.35	.58	.87	
		Swedes	256	6.95	69.49	2.27	1.21	.71	.92
VWO	50	Dutch 1016	35.09	70.18	6.57	2.95	.80	–	
V-2	11	Dutch 1016	8.49	77.18	1.70	1.30	.42	.77	
		Swedes	217	6.40	58.15	2.23	1.47	.57	.86
V-3	11	Dutch 1016	7.64	69.42	2.15	1.36	.60	.87	
		Swedes 99	6.12	55.65	2.69	1.40	.73	.92	

*Spearman-Brown estimate of reliability (KR-20) at k=50

The VWO examination of that year had turned out to be too easy. The reliability (KR-20) of both tests is acceptable. The subtests from the HAVO examination were representative for the total test with respect to their difficulty, as can be seen from the percentage scores (61% and 62%) obtained by the Dutch sample. Of the VWO subtests, however, V-2 with over 77% correct answers, is clearly very much off-target for the intended

population of Dutch students. This could partly explain the disappointing reliability (.42) of this subtest for the target population.

The Swedish groups attain higher scores than the Dutch HAVO students had attained at their examination, but lower than the VWO students on theirs (all differences significant at p <.001). These first results indicate that the ability of the four Swedish groups in their varying numbers of students and in any of the combinations from the streams occurring in the design (cf. Table 1), falls somewhere in between the average ability of the two Dutch levels in this study. Secondly, the results show that the variances of the scores obtained by the Swedish groups are significantly larger (p <.001) than those observed in the Dutch samples on the VWO-subtests, but not on the HAVO subtests (p=.54 for H-4; p=.02 for H-5).

Table 3 reveals the differences in ability between the streams in upper secondary education in Sweden. The first group of columns presents the average results per stream on the Dutch examination subtests. As subjects attempted different numbers of items, the results are expressed as a proportion of all items tried by each subject. The second group of columns presents the average grade on the Swedish Centralt prov (only for the students for whom these results were available).

Table 3
Average results on Dutch examinations (correct proportion of total attempted) and on Swedish tests for Swedish students, per stream.

	Dutch Examination			Swedish Tests		
Stream	n	X(prop)	Stdev	n	Grade	Stdev
Humanities	57	.58	.17	38	3.34	.74
Social Sciences	119	.59	.19	57	3.35	.71
Natural Sciences	131	.73	.17	94	3.71	.92
Engineering	26	.58	.23	21	3.33	.71
Total	333	.65	.19	210	3.51	.84

The results presented in Table 3 show the same kind of ordering of the four streams on the two tests, which means that on average the variable measured by the Dutch subtests is related to the variable measured by the Swedish tests. Also correlations between an end-of-term test and the Dutch subtests which ranged from .47 to .58 (.70 to .84 after correction for attenuation) indicate that the Dutch examinations measure an ability that is related to competencies valued in Swedish education.

All responses of the Swedish students to the items from the Dutch examinations were combined in an IRT analysis with their results, if available, on the Centralt prov, where the observed grades on the latter from 1 through 5 were considered as categories in a single polytomous item. The fit of the first threshold in this item (obtaining a grade of 2

rather than 1) could not be evaluated because there was only one observation in the lowest category. For the other three thresholds (obtaining 3 rather than 2, 4 rather than 3, and 5 rather than 4) the model could not be rejected with probabilities .77, .31, and .92, respectively. The difficulty parameters of the thresholds (-3.433, -1.284, .683, and 2.687) were all estimated at about two logits apart. The equidistance between the thresholds may be a reflection of the use of standard scores in the Swedish grading system. Table 4 provides an overview of the estimated and observed numbers of students in each of the 5 categories. It can be concluded that the Centralt prov seems to measure the same ability as the majority of the items in the Dutch reading comprehension tests.

Table 4
Observed and expected number of students in Centralt prov categories

Category	N-Expected	N-Observed
1	3.15	1
2	20.58	22
3	76.86	79
4	82.32	84
5	27.09	24

4.2 Prediction of (sub)test scores

To investigate the possibility of estimating a score transformation function, which would allow the prediction of subtest difficulty for the Dutch examination populations from item responses observed in Sweden, we used a common item equating design. First item parameters were estimated for the two sets of Dutch examination data separately. Secondly, item parameters were estimated using only the item responses from the Swedish students. The difference between the difficulty of subtests H-4 and H-5 as estimated from the Swedish responses (-.003) was then compared to the difference in difficulty of these subtests as estimated using the responses from the Dutch students (.089). The difference between the two estimates was not significant ($p < .05$). In terms of scores, given the observed score of the Dutch students on subtest H-4 (60.55%), the Swedish data allow the prediction with 95% confidence that the Dutch students will obtain a score between 56.05% and 68.27% correct on subtest H-5. The observed score actually was 62.30 % (cf. Table 2). Similarly the difference between subtests V-2 and V-3 was estimated independently from the Swedish and Dutch responses to be -.041 and -.431, respectively. The observed score for Dutch students on V-3 (69.42%) falls in the 95% confidence interval ($69.14\% < X < 76.05\%$) of the estimate from the Swedish responses.

It can be concluded that relative subtest difficulties for Dutch students at the HAVO and VWO level can be predicted from the results on the subtests as observed from Swedish students of the streams in the Gymnasieskolan involved in this study. For pretesting purposes the confidence intervals are quite acceptable. To equate test forms to be used for testing different cohorts of examinees, however, higher precision is required. This can be achieved first of all by using larger groups. Secondly it must be remembered, that the items in these tests had not been pretested, which was the reason for our study. Using the responses of the Swedish students to the HAVO-items only, the model could not be rejected (p=.11). But the overall fit of the Swedish responses to the VWO-items, of the Dutch responses to both examinations, and the total design including all responses from Swedish and Dutch students on both tests were highly significant (p<.0001). As the model is rejected the data in fact do not allow equating.

4.3 Predicting differences in difficulty of items

Item parameters were estimated independently using the responses of the Dutch examination students and the Swedish try-out groups. The items from the HAVO and VWO examinations were brought onto a common scale using a transformation function calculated from the item parameter estimates from the Swedish responses. The difference in difficulty between the entire HAVO and VWO examinations was estimated at .730.

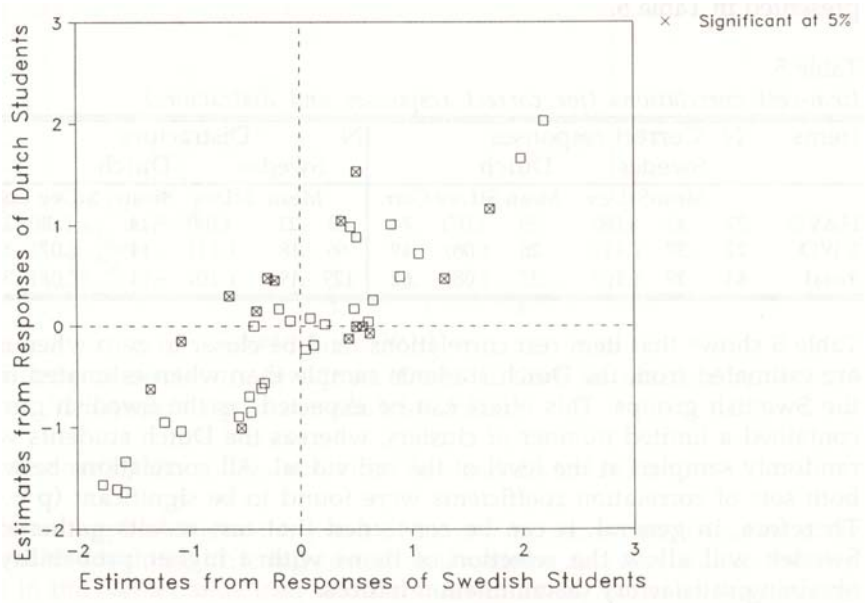


Figure 1
Comparison of item parameters (n=43) estimated independently from Swedish and Dutch students' responses.

Next we evaluated the difference between the item estimates using the responses from the Dutch examinees and the estimates form the try-out in Sweden. Out of the 43 items in the design 15 items fell outside the 95% confidence interval. This high proportion of violations of the model is not unexpected, given that even if only responses from the target examination populations were used, no model fit could be obtained. Within the purpose of this study, however, a statistical test may be less relevant than a visual inspection of the data. Figure l presents the item parameters of the 43 items in the design, estimated independently from the responses of the Swedish and Dutch students. Each box represents one item. Items for which the calibrations differ significantly (p <.05) are marked by a cross. Figure 1 shows that items that are extremely easy or difficult for the Dutch students can readily be detected by using the responses from the Swedish students.

4.4 Prediction of item-rest correlations

For the selection of items from pretests the item-rest correlations both for the correct response (rir) and for the distractors (rdr) are often used as indicators of item quality. It would therefore be useful if some relation existed between these correlations as observed from the try-out in Sweden and as computed from the target population data. To investigate this relation we calculated the mean and standard deviation (using Fisher's z-transformation) of the coefficients estimated from the Dutch and Swedish response sets separately and computed the correlation between both as presented in Table 5.

Table 5
Item-rest correlations (for correct responses and distractors)

Items	N	Correct responses					N	Distractors				
		Swedes		Dutch				Swedes		Dutch		
		Mean	Stev	Mean	StDev	Corr.		Mean	StDev	Mean	StDev	Corr.
HAVO	21	.41	(.09)	.29	(.07)	.74	63	-.21	(.09)	-.14	(.08)	.35
VWO	22	.37	(.11)	.26	(.08)	.49	66	-.18	(.11)	-.14	(.07)	.38
Total	43	.39	(.10)	.27	(.08)	.62	129	-.19	(.10)	-.14	(.08)	.38

Table 5 shows that item-rest correlations tend be closer to zero when they are estimated from the Dutch students sample than when estimated from the Swedish groups. This effect can be expected, as the Swedish groups contained a limited number of clusters, whereas the Dutch students were randomly sampled at the level of the individual. All correlations between both sets of correlation coefficients were found to be significant (p <.05). Therefore, in general, it can be concluded that test results gathered in Sweden will allow the selection of items with a higher probability of obtaining satisfactory discrimination indices.

4.5 Predicting distractor attractiveness

To evaluate predictability of distractor attractiveness we used the item difficulty estimates from the try-out in Sweden and the observed proportions of Swedish students opting for each of the distractors. The difference between the average ability of all Swedish students in the design and the two populations of Dutch students was then used to predict the proportions of Dutch students opting for each of the distractors. As the design comprised 43 items, each with 4 options, the attractiveness of a total of 129 distractors could be estimated. The estimates were compared with the observed proportions of the Dutch students opting for each of the distractors. For 85 distractors the estimate was not significantly different ($p < .05$) from the observed proportion. Figure 2 presents the estimated and observed popularity of the 63 distractors in the HAVO subtests and shows that the practical value of the estimate is greater than the statistical significance test might suggest.

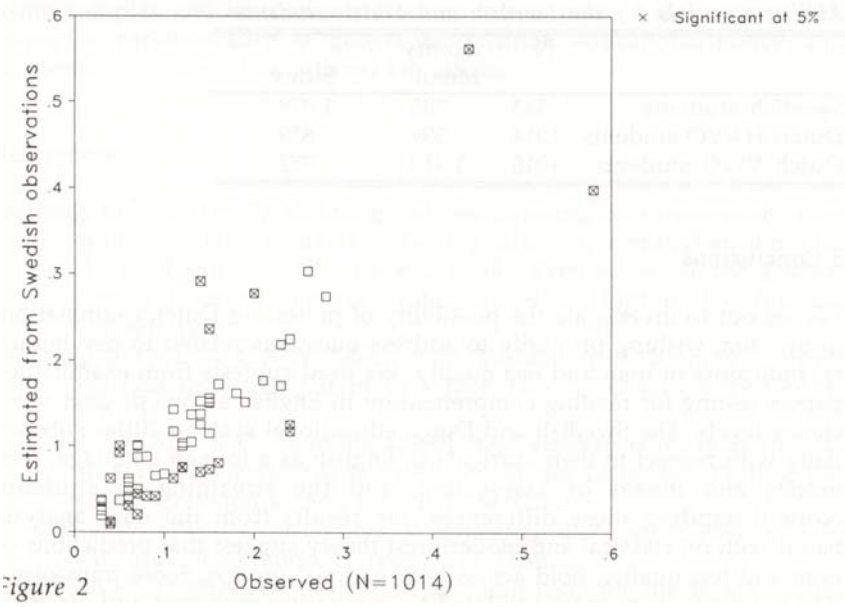


Figure 2

Popularity of distractors (n=63) for Dutch students as observed vs. as estimated from popularity for Swedish students

4.6 Comparing the ability of the Swedish and Dutch students

From an applied-linguistics point of view it is interesting to evaluate the extent to which the differing educational systems and curricula in Sweden and in the Netherlands lead to differences in ability between the students from Sweden and the Netherlands. Table 6 presents the ability estimates for all the Swedish students in the design and for the two populations of Dutch students.

Table 6 reveals that the ability of the average Swedish student in the design is estimated to lie about half way between the ability of the Dutch HAVO and VWO students. This finding seems to suggest that in spite of the substantial differences between the national programs with respect to streaming, curriculum, examinations, etc, the ability level attained by the groups in our study can be predicted from the number of teaching hours and the section of the age cohort. The HAVO and VWO students taken together represent the 35% of the age cohort expected to continue their education at the tertiary level. Together they have had between 550 and 600 hours of instruction and attain an average ability estimated at .855. The Swedish students represent the same section of the age cohort, have had virtually the same number of hours of instruction and attain an average ability level of .705 on the same scale.

Table 6
Ability estimates for the Swedish and Dutch students

	N	Ability	
		Mean	Stdev
Swedish students	333	.705	1.079
Dutch HAVO students	1014	.399	.829
Dutch VWO students 1016	1.310	.772	

5 Conclusions

We set out to investigate the possibility of pretesting Dutch examinations in Sweden, wishing primarily to address questions related to psychometric indicators of item and test quality. We used subtests from examination papers testing for reading comprehension in English at two distinct proficiency levels. The Swedish and Dutch educational systems differ substantially with respect to their curricula in English as a foreign language, their modes and means of assessment, and the streaming of students. Notwithstanding these differences, the results from the data analyses based both on classical and modern test theory suggest that predictions on item and test quality, hold across both national groups. Score transformation functions were cross-validated by comparing predicted and observed subtest difficulty. Within the limits of this study predictions were found to be sufficiently accurate for pretesting purposes. Higher precision would be required for equating proposes, and is likely to be attained by using larger and less clustered samples for the pretests and by eliminating misfitting items. It can be concluded that the quality of Dutch examination papers testing for English reading comprehension could be improved by pretesting in Sweden. For instance, it would have been possible to predict that one of the subtests used in the examination for the higher level students in the Netherlands, was in fact more appropriate for testing the students at the lower level. Furthermore, extremely difficult or easy items

and items with item-rest correlations close to zero would have been detected before administration and could have been excluded from the exams.

The results reported also allow the conclusion that in spite of their differences, the two educational systems seem to lead to comparable outcomes. The same ability was found to underlie the national tests for English as a foreign language in both countries, which agrees with the suggestion that at higher proficiency levels subjects become independent of the curriculum and their proficiency can be assessed on a single dimension (De Jong, 1987; De Jong & Henning, 1990). Moreover, taking into account the sections of the age cohort from both countries involved in our study, the level of proficiency attained is also comparable.

The study has shown the feasibility of cross-national cooperation in developing and monitoring procedures and instruments for standardized objective assessment in the domain of language proficiency. Further research is necessary to investigate whether the results can be generalized using samples and instruments from countries more widely apart with respect to variables such as, geographic position, educational policy, and relationship between the national languages.

References

De Jong, J.H.A.L. (1987). Defining tests for listening comprehension: a response to Dan Douglas's "Testing listening comprehension". In: A. Valdman (ed), *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency. Bloomington*, IN: Indiana University.

De Jong, J.H.A.L. & C.A.W. Glas (1987). Validation of listening comprehension tests using Item Response Theory, Language Testing, 4, 170-194.

De Jong, J.H.A.L. & G. Henning (1990). *Test dimensionality in relation to student proficiency*. Paper presented at the Language Testing Research Colloquium, San Francisco, March 2-5.

Glas, C.A.W. (1989) *Contributions to estimating and testing Rash models*. Doctoral Dissertation, University of Twente.

Glas, C.A.W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, 46, 525-546.

Lindblad, T. (1983). In: J. Van Weeren (Ed.), *Practice and problems in language testing* 5, 11-35. Cito, Arnhem.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47,149-174.

National Swedish Board of Education (n.d.) *Assessment in Swedish Schools*. Stockholm: Author.

Orpet, B.R. (1985). Foreign language teaching in Sweden, *British Journal of Language Teaching*, 23, 1, 37-41.

Oscarson M. (1986) *Native and Non-native Performance on a National Test in English for Swedish Students: A Validation Study.* Report No. 1986:03. Gothenburg, Gothenburg University.

Rasch, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmark Paedagogiske Institut (Chicago: University of Chicago Press, 1980).

Verhelst, N.D. & T.J.H.M. Eggen (1989). *Psychometrische en statistische aspecten van peilingsonderzoek* [Psychometric and statistical aspects of national assessment research]. PPON-Rapport,4. Arnhem: CITO.

Verhelst, N.D. & T.J.H.M. Eggen (1990). *The use of Item Response Theory in the Dutch National Assessment Program*. Paper presented at the 16th International Conference of the International Association for Educational Assessment (IAEA), Maastricht, June 18-22.

Verhelst, N.D., Glas, C.A.W., & H.H.F.M. Verstralen (forthcoming). OPLM: a computer program and manual. Arnhem: CITO.

Verhelst, N.D. & F.H. Kamphuis (1990). Statistiek met ? [Statistics with ?]. *Specialistisch Bulletin*, 77. Arnhem: CITO.

Verhelst, N.D. & H.H.F.M. Verstralen (forthcoming). *The Partial Credit Model with non-sequential solution strategies*. Arnhem: CITO.

Verhelst, N.D., Verstralen, H.H.F.M., & T.J.H.M. Eggen (forthcoming). *Finding starting values for the item parameters and suitable discrimination indices in the One Parameter Logistic Model*. Arnhem: CITO.

Warm, T.A. (1985). *Weighted likelihood estimation of ability in Item Response Theory with tests of finite length*. Technical Report CGITR-85-08. Oklahoma City: U.S. Coast Guard Institute.

THE HEBREW SPEAKING TEST: AN EXAMPLE OF INTERNATIONAL COOPERATION IN TEST DEVELOPMENT AND VALIDATION.

Elana Shohamy, Tel Aviv University
Charles W. Stansfield, Center of Applied Linguistics

1 Introduction

Languages are taught in different contexts and for different purposes. Sometimes a language is taught as a second language so the learners will be able to communicate with the surrounding group that speaks it; at other times a language is taught as a foreign language when the surrounding group does not speak it. It is only natural that teaching methods, materials, curricula and tests will match these different contexts and purposes of language learning. Too often, however, all language learning contexts are treated the same. In the area of test development, for example, it is rarely the case that different test versions are developed and/or used for different language learning contexts. In fact, most tests are developed under the assumption that all test takers have a similar background and learn the language in a similar context and for similar purposes. This situation exists even though there is evidence that the content of the material tested tends to affect the performance of test takers on language tests; furthermore, since tests are not 'content free,' what is familiar to one group of test takers is not necessarily familiar to another group. In terms of test validation as well, validating a test on a group of learners in one context cannot guarantee that it will be valid for a different group of test takers in a different context.

The purpose of this paper is to describe the development and validation of a speaking test that is intended to be used in two different contexts, a foreign language context and a second language context. Two different, yet parallel versions of the test were developed to match these different contexts. Each of the versions was based on content which is known to be familiar to the test takers in each of the contexts. Furthermore, the validation study of the test was also performed with two different groups of test takers, each representing the potential clients of the tests in each of the language learning situations. The development was performed in two parts of the world and close collaboration was facilitated through electronic networking (BITNET).

The test, the development of which will be described in this paper, is a semi-direct oral test, which was developed for two different groups of

learners - language learners of Hebrew in the USA and learners of Hebrew in Israel. In order to accommodate the different settings and contexts where the language is studied and used, two forms of the test were developed for use in Hebrew language schools for learners in Israel, and two forms were developed for use in North America. The first two forms were administered to foreign students at Tel Aviv University and the other two forms were administered to students at Brandeis University and the University of Massachusetts at Amherst.

In this paper the development and validation of the semi-direct Hebrew test (HeSt), its components, phases of development and validation study will be described. The HeST is part of a long term project to develop oral tests for less commonly taught languages where trained oral interview testers are not easily available. It is an oral test which is structured, yet has a variety of communicative features. It is termed a simulated oral proficiency interview (SOPI) since it elicits oral discourse (as opposed to isolated sentences), yet does not involve face to face communication. Although semi-direct tests of English as a foreign language appeard about a decade ago, the Center of Applied Linguistics (CAL) has recently developed a series of semi-direct SOPI tests that employ visual as well as aural stimuli and which use the ACTFL (American Council of Teaching Foreign Languages) guidelines as a basis for scoring (Stansfield, 1989). So far, CAL has developed tests of this kind for Chinese, Portuguese, Hausa, Indonesian and Hebrew (Clark, 1988; Stansfield and Kenyon, 1988; Stansfield et al, 1989; Shohamy et al, 1989). The development of SOPI tests in French and Spanish is currently underway.

2 Background

The first tests of oral proficiency began to appear in the 70's, mostly for testing the speaking proficiency of prospective teachers. These early tests, which can be termed 'pre-communicative', were usually administered in the language laboratory. The type of speaking tasks that the test taker was required to perform were mechanical repetition of words and sentences. Test takers were asked to supply patterned answers to patterned questions and substitution drills (Shohamy and Reves, 1985). In those early days the testing of speaking proficiency on standardized language tests was in itself an innovation, since previous language tests had generally not included any speaking components. Thus testing speaking on standardized tests proved that oral language could be assessed in spite of the difficulties involved in the testing of students on an individual basis.

With the growing emphasis on communicative teaching, language teachers and testers began to claim that earlier types of oral test were artificial and unauthentic since test takers spoke to machines rather than to other human beings. Furthermore, they observed that the language produced was very different from the way human beings normally speak; e.g., the intonation and the length of sentences were not the same as spo-

ken discourse. Consequently, language testers began talking about other types of oral tests which would be more authentic and direct.

Clark (1975) defined the notion of direct language tests. According to Clark, a direct test of speaking proficiency will be one whose testing format and procedure attempts to duplicate as closely as possible the setting and operation of the real-life situation in which the proficiency is normally demonstrated and would therefore involve a test setting where the examinee and one or more human interlocutors engage in a communicative interaction. Testers try to develop tests which will reflect the very types of tasks performed in a natural foreign language context. Such direct oral tests would be those for which an oral language sample is elicited by a live tester who speaks with a test taker. The oral sample is then rated by the tester and/or by another rater with the aid of a rating scale.

The Foreign Service Institute Oral Interview (OD (Clark, 1975; Jones, 1977, 1978; Wilds, 1975) was an example of such a test. It consisted of a face-to-face oral interaction between a tester (the interviewer) and a test taker (the interviewee). The tester asked questions on a variety of topics, and the test taker's responses provided the oral language sample, which was then evaluated by the tester with the aid of a rating scale. The 01 is widely used by US Government agencies belonging to the Interagency Language Roundtable (ILR) and most recently the 01, in its many versions and variations, has come to dominate the American testing scene as the single most important tool for assessing oral language. It gained wide popularity when the American Council on the Teaching of Foreign Languages (ACTFL) adopted it as its official instrument for assessing foreign language proficiency in American secondary schools and universities; it is known today as the 'Oral Proficiency Interview' (OPI). Along with the OPI, special guidelines for rating and evaluation were developed which provided descriptions of different proficiency levels, ranging from novice' through 'intermediate', 'advanced' and 'superior', (each level including a number of sub-levels), for each of the four language skills. The ACTFL Guidelines for speaking proficiency are used to assess the OPI.

However, the new, more widely available, OPI also brought with it some problems. First, there was the problem of content validity which is the degree to which the test represents the specific domain being tested. In speaking tests this means that there is a need to demonstrate that the oral language elicited by the test represents a wide variety of language aspects and oral interaction (Lantolf & Frawley, 1985; Shohamy, 1988; Shohamy & Reves, 1985). Therefore, there is a question,of whether the OPI, which is based mostly on an oral interview, provides a sufficient representation of other oral interactions such as discussions, reports, and conversations which are important components of speaking ability. (Shohamy, Reves, and Bejarano, 1986).

Another problem with oral interview tests is the effect of contextual variables within the testing context on the assessment. Contextual variables such as the tester, the relationship between the tester and test taker, their personalities, their genders, the purpose of the interaction, the topic, the content, and the setting may all have an effect on the test taker's

oral performance. In fact, in a number of studies (Shohamy, 1988) it has been shown that contextual variables, such as the tester, the type of interaction, the time of testing, etc., do affect the oral language output. There is a need, therefore, to control some of those variables by conducting oral tests in a more uniformed, structured and standardized way. This will ensure the reliability and validity of such oral tests without compromising their communicative features.

Another problem with the OPI is the need for a 'live' tester to be available in order to conduct the test; this can be a problem for the uncommonly taught languages as trained testers are not always available to conduct the interview type tests. The SOPI attempts to remedy one of these problems. In this test, test takers respond to recorded authentic tasks which require them to produce discoursal reactions. The test is more uniform than oral interviews since the testing context is controlled by holding the tester variable constant and by requiring all test takers to perform similar oral tasks. At the same time these tests involve a variety of communicative characteristics that are in line with the current trend of the communicative approach to language. Moreover, they also elicit a wider range of oral interactions and discourse strategies, thus increasing content validity (Stansfield, 1990). From a practical point of view, as well, the SOPI offers considerable advantages as it can be administered by any trained interviewer, whether teacher, aide, or language lab technician. This may be especially useful in locations where a trained interviewer is not available; in the foreign language context, it can often occur that trained raters are not commonly available.

3 The Hebrew Speaking Test (HeST)

3.1 The Components of the Test

The HeST is similar in its structure to the SOPI tests developed by the Center of Applied Linguistics for other uncommonly taught languages: It consists of six parts:

a) *Personal Conversation,* in which the examinee hears ten questions about family, education, hobbies, etc., in Hebrew and is expected to respond to each one in order. This part serves as a warm-up phase.
b) *Giving Directions,* in which the examinee is shown a pictorial map in a test booklet and is instructed to give directions according to a map. In this task and in all subsequent sections, the test tasks are in English.
c) *Detailed Description,* in which the test taker is shown a drawing of a particular scene or place including a variety of objects and actions, and is expected to describe the picture in detail.
d) *Picture Sequence,* in which the test taker is instructed to speak in a narrative fashion about three sets of four pictures in sequence, each set of pictures requires the use of a different tense - past, present and future.

e) *Topical Discourse,* in which the examinee is instructed to talk about five topics, each involving specialized content and varying in difficulty and discourse strategies.
f) *Situations,* in which the test taker is asked to respond to five authentic situations in which a specified audience and communicative task are identified. The situations test the ability to handle interactive situations through simulated role-playing tasks – tasks involving the use of appropriate speech acts such as requesting, complaining apologizing and giving a formal speech.

All the tasks are read aloud in English (L-1) on the tape and are also written in English in the text booklet, so that any difficulties test takers may have in comprehending the questions will not affect the test taker's performance in speaking. The use of Parts b, c, and d are based on pictures,
which appear in a test booklet which the test taker is using. The test taker's responses are recorded on another tape-cassette recorder.

Two versions of the Hebrew Speaking Test (HeST) were developed. One version, the U.S. version, is specifically designed for test takers who may not be familiar with Israeli culture, whereas the second, the Israeli version, is intended for students who have been to Israel and are familiar with its culture. Two parallel forms were developed for each of these versions - U.S. version, Forms A and *B;* Israeli version, Forms A and *B.*

3.2 The Phases of Test Development

Most of the day-to-day test development work was conducted at Tel Aviv University, Israel, by two specialists in language testing, through regular, almost daily, consultation with test developers at the Center for Applied Linguistics in Washington, D.C. A test development committee was formed which included, in addition to the above, one experienced teacher of Hebrew as a Foreign Language at the university level and an experienced material developer. The committee met on a regular basis from November 1988 to January 1989 to develop the specific items for the test. These items were based on the question types used in the SOPI test of Chinese and Portuguese.

Communication between Tel Aviv and Washington was facilitated by the use of electronic mail (BITNET). For example, after items were drafted in Israel, they were sent via electronic mail to CAL for review and comment, and then returned to Israel for further revision. Answers to urgent questions and clarifications were available within hours despite the distance between the two centers. Electronic mail allowed the continuous and efficient exchange of information throughout the project.

3.3 The Trialing

Each of these four forms was trialed on eight to nine students representing three different levels of proficiency. Forms A and *B* of the U.S. version were trialed in the USA; Forms A and *B* of the Israeli version in Israel. The purpose of the trial was to ensure that the questions were clear,

understandable and working as intended, as well as to check the appropriateness of the pause times allotted on the tape for examinee responses. The subjects took the test on an individual basis using two tape-recorders. Upon completion of the test, the examinees responded to a detailed questionnaire about it. When possible, they were also questioned about the test in person. In most cases the students were observed while taking the test by a Hebrew speaking member of the test committee who took notes on the students' performance and also responded to a questionnaire about the test. Based on the students' questionnaires and comments, modifications in the questions were inserted. In most cases they involved clarification of ambiguous items in the tasks and minor revisions in the pictures. The original pauses were adjusted - they were lengthened or shortened in various items. As a result of the trialing, it was decided to prepare two versions of the Hebrew warm-up conversation, one for female test takers and one for male test-takers. This decision did not alter the wording of the scripts significantly.

4 The validation study

Two separate validation studies were carried out, one for the Forms of the Israeli version, those intended for learners of Hebrew in Israel, and the other for the Forms of the U.S. version, those intended for learners in the USA. The validation was carried out on the two versions of each form. The validation study sought to answer the following questions:

a) Can each test version, which involves a spoken response in Hebrew, be scored reliably by different raters?
b) Are the two forms (A and B) of each version interchangeable, i.e., do they produce similar examinee results *independently of* the particular form administered?
c) Do the recorded responses produce the same score as a regular live interview for any given examinee?

To answer these questions, a research design was prepared involving 40 subjects. In the two parallel validation studies the Form A and B of the U.S. version were validated in the USA on 20 students learning Hebrew, while the validation of Form A and B of the Israeli version was conducted in Israel with 20 English speaking students studying Hebrew in Israel. Each subject was administered two versions of the appropriate test form of the HeST and the Oral Proficiency Interview (OPI). The design controlled for order of administration, with half of the subjects receiving the form A first and form B second; the other half in reverse order. In all cases the OPI was administered before the administration of the HeST. The design also attempted to control for proficiency levels; students from three different class levels were selected for participation.

The subjects were predominantly undergraduate students who had completed a year or more of Hebrew study at their respective universities.

Each subject received a small honorarium for participating in the study. Some of the subjects took the HeST in the language lab and others on an individual basis where their responses were tape-recorded. The OPIs were administered by testers who have had experience in administering oral interviews in a variety of settings. The OPIs were administered individually and tape recorded, to be rated at a later date by the two raters.

Four raters, two in the USA and two in Israel, rated the speech samples. The Israeli raters were experienced Hebrew teachers who received intensive training in using the ACTFL guidelines prior to the ratings of the speech samples of the validation study. The raters in the USA are trained ACTFL raters. The ratings of all the tapes were done independently, anonymously and in random order; however, all the OPIs were rated before rating the HeST tapes.

Rating of both the live interview and the tape based SOPI tests was done on the ACTFL scale with weights assigned as in Figure 1.

Figure 1
Rating Categories and Weights

novice-low	0.2
novice-mid	0.5
novice-high	0.8
intermediate-low	1.0
intermediate-mid	1.5
intermediate-high	1.8
advanced	2.0
advanced-plus	2.8
superior	3.0
superior-plus 3.8	

The scores of the two raters on the live interview and on each of the HeST forms from both the U.S. version and the Israeli version validation studies were sent to CAL via electronic mail for analysis.

5 Results

To answer the question regarding the degree to which the HeST can be scored reliably, inter-rater reliability (Pearson product moment correlations) was assessed between the ratings assigned by two independent raters for each of the SOPI forms and for the live interview. These are shown in Table 1.

Table 1

Interrater Reliabilities for Forms A and B of U.S. and Israeli Versions of HeST and for the Oral Interview

Test Version/Form	Correlation (r)
U.S. version (N=20)	
Form A	.92
Form B	.92
Oral Interview	.98
Israeli version (N=20)	
Form A	.93
Form B	.97
Oral Interview	.99

Results showed that the interrater reliabilities were uniformly high across the four SOPI forms and the live interview. This suggests that the HeST elicits a sample of speech as ratable as the live interview.

In order to answer question two, the degree to which the two forms of each version are interchangeable, i.e. parallel forms, correlations were calculated for the same subject taking the two different forms of the either the U.S, version or the Israeli version of the HeST and these were rated by the same rater. Table 2 shows the correlations indicating parallel form reliability.

Table 2

Parallel Form Reliabilities (Same Rater)

Version	Rater I	Rater II
U.S. version		
Form A*Form B (N=20)	.99	.93
Israeli version		
Form A*Form B (N=20)	.94	.94

The statistics indicate that the parallel form reliability of the HeST is very high.

Regarding the concurrent validity of the HeST with the OPI correlations between the live interview and the different versions of the test were calculated, results of which are shown in Table 3.

As indicated by the high correlations obtained, the both versions of the HeST seem to be measuring oral language proficiency to a similar degree as the live interview and therefore can be substituted as an alternative but equally valid form of assessment.

Table 3
Correlations of the HeST with the Oral Interview

All Matched Pairs	Correlation (r)
Interview*U.S. version (N=20)	.93
Interview*Israeli version (N=20)	.89

In addition to the above, the test taker's attitudes towards each of the test versions was assessed. The questionnaire elicited information on various aspects of their experience with and opinions about both types of testing procedures. The questionnaire was given to the subjects directly after they completed the SOPI tests. The results of the questionnaires indicated that though the subjects were very positive about the content, technical quality and ability of the HeST to probe their speaking ability, the unfamiliar mode of administration and the 'unnaturalness' of speaking to a machine caused greater nervousness than the live interview for most of the test takers. Thus, the majority of the subjects said that they preferred the live interview to the taped test.

6 Conclusions

The above results indicate that both versions of the new semi-direct Hebrew Speaking Test offer good and efficient alternatives to the OPI. SOPI tests that control some contextual variables may have some advantages over testing procedures where a test taker's language may be affected by certain variables in the testing context. SOPI tests may even be preferable to oral interviews in situations where trained testers are not available Another advantage to this type of test is its content validity. By requiring the test takers to perform a variety of speech acts and discourse strategies such as describing; giving instructions, narrating etc., the HeST can provide a broader representation of the test takers' oral language proficiency thereby resulting in a more valid assessment. In terms of its psychometric properties, the HeST has high inter-rater reliability and high concurrent validity with the OPI, indicating that oral proficiency can be tested efficiently and reliably. A major advantage of the HeST is the fact that the test's items and tasks are based on content which is familiar to the test taker in the specific contexts for which s/he is learning the language. The very favorable validation study results in each of the Hebrew learning contexts indicate that these both versions are appropriate for learners in the specific milieu where the language is learned and used. Another advantage of the test is that it can be administered simultaneously to a group of test takers in a single administrator whereas the OPI must be individually administered. In that respect the SOPI is preferable especially when trained interviewers are not available. The SOPI also allows more standardization as the questions in each format are the same as opposed

to the oral interview where questions vary from one interviewer to another.

Given the extremely high correlations between the two types of tests and the positive responses to the taped test quality it appears that the taped test may confidently be used as an alternative to the live interview. It is expected that test takers who are more prepared for this type of test may find the testing mode less threatening than the subjects participating in the validation study who went to the test without any special advance preparation.

However, in spite of the high correlations between the OPI and the HeST there are still a number of questions about semi-direct tests which require further and more in-depth research. Although both test formats correlate highly, there is a need to study each format in a more thorough manner in order to determine the specific aspects of oral proficiency that each test format taps. In order to examine this issue, an analysis of the language obtained through each of the formats will be performed. This analysis will provide evidence of content validity of each format, i.e. the specific aspects of oral language such as range and quantity of lexical items, register shifts, speech acts, etc., elicited by each of the formats. This will provide insight as to whether the language output obtained through taped semi-direct tests is similar to the language obtained from more authentic/direct face to face interactions such as the OPI.

The development and validation of the HeST shows that it is possible to combine human and material resources on different continents and work cooperatively on international projects without a loss in efficiency. This was possible because of the daily use of rapid communications, especially BITNET. Courier mail and FAX was also used on occasion. Clearly however, without BITNET the project could not have been completed on schedule, and therefore, international cooperation among language testing specialists would never have been attempted. BITNET was designed to facilitate networking among academics in diverse locations. However, it has far greater potential. In this case, BITNET facilitated the internationalization of a project that otherwise would have been only national in scope. We plan to continue to utilize BITNET for other cooperative projects among ourselves and others.

Acknowledgement

This project was funded by the U.S. Department of Education via grant number 6008740397 to the Center of Applied Linguistics, Washington D.C. The HeST is administered by the Center of Applied Linguistics, 1118 22nd St. Washington D.C. 20037, phone 202-429-9292.

References

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

Clark, J.L.D. *(1975).* Theoretical and technical considerations in oral proficiency testing. In R.L. Jones and B. Spolsky (eds). *Testing Language Proficiency.* Arlington, VA: Center of Applied Linguistics, *10-28.*

Clark, J. L. D. *(1988).* Validation of a tape-mediated ACTFL/ILR-scale based test of Chinese speaking proficiency, *Language Testing 5, 187-98.*

Jones, R. A. *(1977).* Testing: a vital connection. In June Phillips, (ed) *The Language Connection: From the Classroom to the World.* The ACTFL Review of Foreign Language Education. Skokie, IL: National Textbook *Co.,9: 237-265*

Jones, R. L. *(1978).* Interview techniques and scoring criteria at higher proficiency levels. In John Clark, (ed) *Direct Testing Speaking Proficiency: Theory and Application.* Princeton, N.J.: Educational Testing Service, *89-102.*

Lantolf, J., and Frawley W. *(1985).* Oral proficiency testing: A critical analysis. *Modern Language Journal 69, 337-343.*

Morrow, K. *(1977). Techniques of Evaluation for Notional Syllabus.* Reading: Center for Applied Language Studies, University of Reading.

Shohamy E. *(1988).* A proposed framework for testing the oral language of second/foreign language learners. *Studies in Second Language Acquisition 10, 165-79.*

Shohamy E., Reves, T., and Bejarano, Y. *(1986).* Introducing a new comprehensive test of oral proficiency. *English Language Teaching journal 40, 212-220.*

Shohamy, E. and Reves, T. *(1985).* Authentic language tests: Where from and where to? *Language Testing 2, 48-59.*

Shohamy, E., Gordon, C. Kenyon, D.M., & Stansfield, C.W. *(1989).* The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. *Bulletin of Hebrew Higher Education 4, 1, 4-9.*

Stansfield, C.W. and Kenyon, D.M. *(1988). Development of the Portuguese Speaking Test.* (Year one project report on development of semi-direct tests of oral proficiency in Hausa, Hebrew, Indonesian and Portuguese.) Alexandria, VA: ERIC Document Reproduction Service, ED *296 586.*

Stansfield, CW. *(1989). Simulated Oral Proficiency Interviews.* ERIC Digest. Washington, DC: ERIC Clearinghouse for Languages and Linguistics and the Center for Applied Linguistics.

Stansfield, C. W. Kenyon, D.M., Paiva, R., Doyle, F. and Ulsh, 1. *(1990).* The development and validation of the Portuguese Speaking Test. *Hispania, 73,3.*

Stansfield, C.W. (1990). A *comparative analysis of simulated and direct oral proficiency interviews.* Plenary presentation at the Regional English Language Center Conference, Singapore.

Wilds, C. (1975). The oral interview test. In Randall Jones and Bernard Spolsky (eds) *Testing Language Proficiency.* Arlington, VA: Center for Applied Linguistics, 29-44.

EUROCERT: AN INTERNATIONAL STANDARD FOR CERTIFICATION OF LANGUAGE PROFICIENCY

Alex Olde Kalter, European Agency of Educational Testing Service
Paul W.J.E. Vossen, University of Nijmegen

1 Introduction

The countries of the European Community have agreed to abolish national borders between the member countries in 1992. Many changes are expected to result from this resolution. It will affect the mobility of the public and implies the removal of economic restrictions in the commercial field. European citizens will seek employment or study abroad and companies will experience greater opportunity to expand internationally. This development is bound to increase the need for all citizens to learn languages other than their own and certification programs for the European languages will be needed. In view of its current stature both in Europe and in the rest of the world, this will apply in particular for English.

A certification program for English would not only have to be accepted by universities in English-speaking countries as a qualification for entrance, but would also have to be acknowledged in trade and industry. The program will have to comprise a battery of tests testing for various aspects and domains of foreign language proficiency, Two quite different courses of action can be taken to establish such a battery: (a) developing a completely new battery of tests, or (b) giving meaning in this new context to already existing tests.

The program EUROCERT has chosen the second course of action. It makes use of the ETS (Educational Testing Service) tests TOEFL (Test of English as a Foreign Language), TWE (Test of Written English) and TSE (Test of Spoken English) in a context for which they were not originally intended. This three-test combination enables the measuring of the four skills: reading, writing, speaking and listening. The basic purpose of EUROCERT is to provide an international certificate for English. The EUROCERT program is jointly developed and sponsored by Educational Testing Service (ETS) and the Dutch Institute for Educational Measurement (CITO).

In this paper the composition of EUROCERT will be introduced. Subsequently, an attempt will be made to start inquiries into the "appropriateness" of the TSE in the light of EUROCERT. We will devote

special attention to the development, background and use of the oral component (the TSE). A comparison will be made of the functioning of this test as a component of EUROCERT with its functioning in its regular international administration. Differences between the samples may occur in that candidates may come better prepared when they sit for EUROCERT, knowing that if they do not meet the requirements, they get no certificate and have, so to say, achieved nothing. Regular candidates, on the other hand, may always find that some university will accept the scores they have achieved, there are no minimum requirements. For our study we made use of two kinds of data:

a. an inquiry into the acceptance of test results by score users, in particular employers;
b. an analysis of scores of regular candidates and candidates who took the test under the flag of EUROCERT.

2 Background and development of the TSE

The TOEFL was developed in the sixties and designed to measure the receptive skills, reading comprehension and listening comprehension, and aimed to provide also a predictive measure of students' productive command of written language, Its major purpose is to evaluate the English proficiency of people whose native language is not English. It has been developed in order to measure the English proficiency of international students who wish to study in the United States. This is still the primary function of the test. The TCEFL consists of three sections (for further details, see Educational Testing Service, 1990a):

Section 1	Listening Comprehension, measuring the ability to understand spoken English.
Section 2	Structure and Written Expression, measuring the extent to which an examinee has command of several important structural and grammatical points in standard written English.
Section 3	Vocabulary and Reading Comprehension, testing the ability to understand meanings and uses of words and the ability to understand a variety of reading materials.

In the last decade doubts have arisen about the predictive claims of Section 2 and the need for separate measures of the productive skills speaking and writing was felt. Though intercorrelations between productive and receptive skills reveal a fairly strong relationship, caution is required when statements about speaking and writing proficiency are made at the level of the individual, as has been concluded by Clark and Swinton (1979).

Therefore, two additional tests were developed, the Test of Written English (TWE) and the Test of Spoken English (TSE). Research studies (Hale and Hinofotis, 1981; Angelis, 1982; Kane, 1983; Carlson, Bridgeman,

Camp and Waanders, 1985) provided the foundation for the development of the Test of Written English. TWE topics are based on the types of writing tasks identified in the Bridgeman and Carlson (1983) study (Educational Testing Service, 1989). And, based on the findings of the validation study, a single holistic score is reported for the TWE. This score is derived from a criterion-referenced scoring guide that encompasses relevant aspects of communicative competence. The TWE is offered as an additional part of TOEFL at 4 out of the 12 administrations a year (for further details on the TWE, see Educational Testing Service, 1989).

The primary purpose of the TSE is to evaluate the English speaking proficiency of persons whose native language is not English. Though the TSE was initially intended for use in the United States and Canada in academic circles and later in health-related professions (Powers & Stansfield, 1982), we see already today that it has also become important in the European context of trade and industry. When applying for a job for instance, TSE scores are often used by the applicant to give an indication of his/her level of oral proficiency in English. Like the TOEFL and the TWE, the TSE is not targeted on particular situations of language use, thus permitting the examinee to demonstrate general language proficiency.

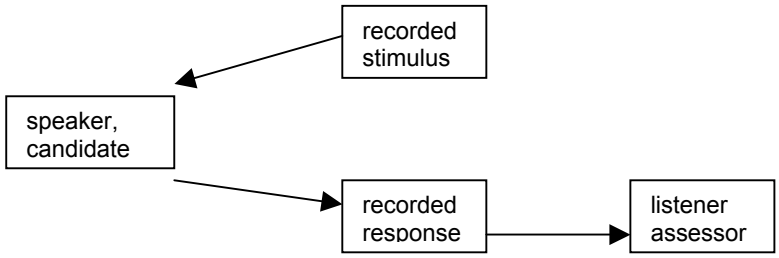
In the last few decades "communicative competence" has become an increasingly important issue in the fields of language teaching and testing. It is now generally recognized that learners have a need for practical foreign language command as opposed to the metalinguistic knowledge valued in former years. People are more and more interested in how to use a foreign language in normal everyday-life situations (see, e.g., Van Els et al., 1984). Canale (1981) pleads for more practicality in the evaluation of communication because of the shift in emphasis from language form to language use. He points out two main ways for achieving this goal. Firstly, by reducing the time and the cost involved in test administration and scoring, and secondly, by developing tests that address core aspects of communication and from which the results would be maximally generalizable.

By many authors the evaluation of oral proficiency is considered the most difficult part of testing foreign language skills (e.g. Harris, 1969; Clark, 1972; Heaton, 1975). The biggest problems here lie in general with the administration and scoring of the tests.

The two most important operationalizations of tests of oral proficiency are direct testing and semi-direct testing, or in more concrete words, the oral interview and the recorded test, also called oral proficiency interview (OPI) and simulated oral proficiency interview (SOPI) (Stansfield, 1989). According to Stansfield (1989), a semi-direct test was defined by Clark (1979) as a test that elicits speech by means of tape recordings, printed test booklets, or other non-human elicitation procedures. Semi-direct testing offers better opportunities for standardization and better guarantees for objectivity, both important characteristics for tests that are administered all over the world. Objectivity can only be accomplished if the testing takes places under standardized conditions. Stan-

dardization increases the reliability of test results and without it, performance on a test may be influenced by differences in procedure and thus reduce the quality of interpretations of test results (American Educational Research Association, et al., 1985). Figure 1 depicts the procedure of semidirect oral proficiency testing.

Figure 1
Schematic representation of semi-direct oral proficiency testing



Recorded tests have many advantages but also some disadvantages. The most important advantages, according to Underhill (1987), are that, in case a language laboratory is used, many people can be tested at the same time. Another advantage is that the marking can occur when the assessor feels comfortable and at a time s/he chooses. Also the fact that each examinee is given exactly the same questions makes distinctions in proficiency more salient to the rater. Probably the most important advantage of recorded tests is that they allow a standardized operation and can be used at times and places where it is difficult to test the examinees one by one individually.

A disadvantage mentioned by Underhill (1987) is that a recorded test lacks authenticity and does not show the assessor everything that an interviewer, in for example, the FSI oral proficiency interview, may see (things like gesture and facial expression). This may jeopardize the validity of the test. Though direct oral tests are often considered more authentic than semi-direct tests, also direct oral tests have problems with validity. The fact that attempts are made to elicit certain structures or lexical items to compensate for the random use of language makes the direct oral test often less natural. The best way generally is to establish a compromise between naturalness and efficiency (Madsen & Jones, 1981).

The growing interest in aspects of communicative competence and the need for measures of oral proficiency testing led to the development of the TSE, the Test of Spoken English, by Educational Testing Service in 1978 (for more details, see Educational Testing Service, 1990b). The TSE initially comprised two parallel forms, each including eleven question types in a separate test book and on a tape recording. The question types had been identified for incorporation into an experimental test after examining a variety of question types used in previously developed 1. semi-direct" speaking tests and a series of project staff meetings. This

experimental test was individually administered to twelve examinees. The primary goal was to obtain information on the overall testing procedure rather than to obtain extensive statistical data. Examinees had previously taken the TOEFL and participated in a taped face-to-face oral proficiency interview rated according to the standard Foreign Service Institute rating procedure.

Based on the information obtained, judgements were made concerning each of the eleven item types and a new edition of the test was prepared, again consisting of two new parallel forms containing those item types judged to be effective. Both forms were administered to 155 examinees, who also took the TOEFL, completed a questionnaire and participated in a FSI-type interview. Performance on each test item was compared with performance on the TOEFL and on the oral interview. Within each item type, those items showing the greater correlations with the FSI criterion scores were identified and recommended for inclusion in a prototype TSE. The specific items within each item type were selected with the goal of maintaining the highest possible correlation with the TOEFL score. A pilot testing under actual operational conditions determined the practicality of a speaking test administration at TOEFL test centers and identified operational aspects such as procedural modifications required and administration procedures.

A major consideration in developing a measure of speaking proficiency was that it had to be amenable to standardized administration at TOEFL test centers worldwide. This factor immediately eliminated faceto-face interviewing which would involve direct conversation with the examinee and would require a thoroughly trained native or near-native speaker of English as the interviewer. Providing the necessary training in interviewing techniques on a worldwide basis was considered too costly and impractical. Another factor addressed during the development of the TSE was the linguistic content of the test. Because the test would be administered in many countries, it would have to be appropriate for all examinees regardless of native language or native culture. A third factor in test design considerations was the need to elicit evidence of general speaking proficiency rather than proficiency in a particular language-use situation. Because the instrument might be used to determine examinee suitability in a large number of contexts, the test could not use item formats or individual questions that would require extensive familiarity with a particular subject matter area or employment context.

The TSE consists of seven sections, each involving a particular speech activity. The first section is an unscored "warm-up" in which the examinee responds orally to a short series of biographical questions spoken on the test tape. In the second section the examinee reads aloud a printed passage and is told that scoring will be based on pronunciation and overall clarity of speech. In the third section the candidate sees ten partial sentences and is asked to complete them orally in a meaningful and grammatically correct way. The fourth section consists of a number of drawings or pictures that "tell a continuous story". The examinee is asked to tell the story in as much detail as possible. In section five a single line

drawing is presented and the examinee has to answer a series of spoken questions about the picture's content. Section six consists of a series of spoken questions to which the examinee is expected to give somewhat lengthy responses. In the seventh and final section the examinee is presented with a printed schedule which he/she has to describe as if addressing a group of students or colleagues.

The test takes about twenty minutes, is given at TOEFL test centers throughout the world and can be administered to individuals by means of cassette or reel-to-reel tape recorders or to a group, using a multiple recording facility such as a language laboratory. Each section of the TSE has a time limit, the beginning and end of which are indicated by the voice on the tape.

The TSE makes use of divers elicitation techniques. There are several reasons for using more than one technique in a single test. Firstly, it is more authentic to use a variety of techniques. Secondly, since different people are good at different tasks, it is quite normal to make use of a variety of techniques. Another reason is that such a combination is necessary in order to present an overall picture of oral proficiency. All of them are factors that contribute to the objectivity of the test (Underhill, 1987).

Performance on the TSE is evaluated by two raters, randomly assigned from a pool of raters who have a background in language teaching and testing and have attended a rater training workshop at ETS or CITO. Raters evaluate examinees' performance along four linguistic dimensions. Three of these, grammar (GRAM), fluency (FLUE), and pronunciation (PRON), are considered diagnostic scores; the fourth dimension, comprehensibility, (COMP) is considered to be integrative. In the following tables we will denote the linguistic dimensions by their abbreviations, adding the section number if relevant. Thus PRON2 will be used to refer to the dimension of pronunciation as it is operationalized in Section 2 of the TSE.

Each recording of a candidate's responses obtains two sets of ratings. The number of sections and items composing each dimension is shown in Table 1. Each of the linguistic dimensions is rated on a four-category scale, with scores ranging from 0 (minimum) to 3 (maximum). For each dimension, section scores are computed by averaging over all items in the section. An overall score on each of the four dimensions is obtained by averaging across the section scores. For example, the overall score for grammar is the average of GRAM3 and GRAM5. The result is a set of four overall scores for each examinee from two raters. For score reporting purposes, the two sets are averaged. If the two raters differ by more than .95 at the overall score level on any one of the linguistic dimensions, a third rater is brought in. Final scores for tapes requiring third ratings are based on a resolution of the differences among the three scores.

Table 1
Sections that contribute to TSE scores and number of items per section.

Section	k	Linguistic Dimension			
		PRON	GRAM	FLUE	COMP
2	1	PRON2		FLUE2	COMP2
3	10		GRAM3		COMP3
4	1	PRON4		FLUE4	COMP4
5	4	PRON5	GRAM5	FLUE5	COMP5
6	3	PRON6		FLUE6	COMP6
71		PRON7		FLUE7	COMP7
OVERALL*		PRON	GRAM	FLUE	COMP

* The Overall score is the average of the section scores.

One of the TSE's most important features is that it is a recorded test. This ensures identical administration procedures all over the world. Furthermore, this also seems to contribute to highly consistent interpretations of test results supported by the Table 2, indicating interrater reliability of TSE scores (Clark & Swinton, 1980).

Table 2
Interrater reliability of TSE scores (n=134)

TSE Score	Number of Raters per Score		
	1	2	3
COMP	.79	.88	.92
PRON	.77	.87	.91
GRAM	.85	.92	.94
FLUE	.79	.88	.92

The coefficients in Table 2 indicate that the consistency of TSE scores across raters is quite high. The interrater reliability of the overall comprehensibility score for a single rater is .79. This coefficient represents the actual correlation between the scores assigned by two separate raters who scored the same 134 examinee tapes. When an examinee's score is based on the average of two ratings, as is currently the case unless scores are abnormally discrepant, the interrater reliability of the TSE overall comprehensibility score is .88. That is, if tests already scored by two raters were rescored by two different raters, the correlation between the scores would be .88. Three raters are used whenever there is a substantial discrepancy between the ratings assigned by the first two raters to the overall score or any of the diagnostic subscores. Current experience indicates that a third rater is needed about 6 percent of the time. In such cases, the estimated interrater reliability of the overall comprehensibility

score is .92. That is, if tests scored by three raters were rescored by three different raters, the correlation between the two score averages would be .92. Because most TSE scores are based on the average of two ratings, the coefficient .88 can be considered a fair, if not a slightly conservative, estimate of the true interrater reliability of TSE scores.

Together with an alpha coefficient of .91, these figures support the statement that the TSE is a quite reliable test which is also shown by Table 3. It depicts the relationship between the four TSE scores, based on a sample of 134 teaching assistants at nine universities (Clark and Swinton, 1980). The figures indicate that the overall comprehensibility rating is more closely related to pronunciation and fluency (r=.93 and .91, respectively) than to grammar (r=.84). It is important that due to its being a semi-direct test of oral proficiency, the validity of the TSE is kept in mind. Construct validity is revealed by data that show that while TSE scores are closely related, each provides some independent information about examinee proficiency. This is especially true in the case of the TSE grammar subscore, which shows the lowest correlations with the other measures.

Table 3

Intercorrelations among TSE scores (n=134)

	COMP	PRON	GRAM	FLUE
COMP	1.00			
PRON	.93	1.00		
GRAM	.84	.79	1.00	
FLUE	.91	.88	.82	1.00

Another indication of the TSE's construct can be gathered from the correlations between TSE scores and TOEFL scores presented in Table 4. TSE subscores other than grammar show only moderate correlations with TOEFL scores. This indicates that the TSE represents a substantial improvement over TOEFL in the prediction of oral language proficiency. Not surprisingly, the TSE grammar score is more closely associated with the TOEFL total than are the other three TSE scores. This suggests that while the score on grammar is the least predictive of overall comprehensibility within the TSE (cf. Table 3), it is also the most generally relevant subskill across the different modes or channels of language use purportedly measured by the TOEFL. Or to put it in another way, correct use of grammar from Tables 3 and 4 seems to come closest to a kind of general linguistic proficiency.

Table 4
Correlations between TSE component scores and TOEFL total score

TSE Component	TOEFL Total
Pron	.56
Gram	.70
Flue	.60
Comp	.57

For more than three decades the principal test of spoken language has been the oral proficiency interview developed by the Foreign Service Institute of the United States Department of State. (Wilds, 1975; Sollenberger, 1978). It consists of a structured conversation of about fifteen to twenty-five minutes between the examinee and a trained interviewer who is either a native or a near-native speaker of the test language. Performance on the interview is evaluated on a scale ranging from 0 to 5, with level 0 representing no functional ability in the language, and level 5 representing proficiency indistinguishable from that of the educated native speaker. In addition to these five proficiency levels, the FSI scale also assigns a plus to an interviewee who substantially exceeds the requirements for one level and fulfills most, but not all, of the requirements for the next level. Thus, it is possible to obtain FSI scores of 0, 0+, 1, 1+, 2, 2 +, ...etc.

In order to determine the relationship between scores on the FSI interview and scores on the Test of Spoken English, both tests were administered to sixty foreign teaching assistants (TA's) at several large state universities (Clark and Swinton, 1980). In addition, TOEFL scores for thirty-one of the TA's were obtained from student records. Table 5 shows the correlations between the TSE and the FSI oral interview. The fairly high correlations reported in Table 5 are indicative of the concurrent validity of the TSE.

Table 5
Relationship of TSE Scores to FSI Ratings.

TSE Score	FSI Rating
PRON	.77
GRAM	.73
FLUE	.76
COMP	.76

The FSI oral proficiency interview is regarded as a criterion-referenced test. The TSE ratings can also be related to criteria (represented by the scoring scales). Therefore, the TSE can be regarded as criterion-referenced. Its goal is to determine the relative position of a candidate with

respect to a particular continuum of possible scores which can be expressed as a (series of) statement(s) about the skills or subskills which an examinee has or has not completely mastered.

In particular when one is working with a criterion-referenced test, it is extremely valuable to be able to make reliable interpretations of test results. A score is not compared to other scores to determine whether a candidate has passed or failed the test, but weighed against requirements, so even the smallest variation could be fatal. It could mean that a student will not be admitted to a university or that, for example, a EUROCERT candidate does not pass the cut-off scores and consequently will not receive a certificate. Table 6 may be useful in understanding the relationship between a TSE score of overall comprehensibility and overall oral proficiency. Based on a limited sample (n=60), it depicts the relationship between scores on the FSI oral proficiency interview and performance on the TSE (Clark and Swinton, 1980).

Table 6
Approximate Relationship between TSE Overall Comprehensibility Scores and FS1 Oral Proficiency Levels

FSI Level	TSE Score
1 +	150-160
2	170-190
2+	200-210
3	220-240
3+	250-260

The data in Table 6 indicate that the TSE is difficult for persons of limited oral proficiency, and that it discriminates well among speakers of FSI levels 2 to 3+. As such it should not be expected to distinguish between native and very advanced nonnative speakers of English. This is called a "ceiling-effect"; a quite common feature with most oral tests (Madsen & Jones, 1981).

3 Method

In the study conducted, data were first obtained from a survey. Because concrete feedback about candidates' experiences with respect to the EUROCERT certificate was not available, a questionnaire was sent in October 1989 to all candidates who had received a EUROCERT certificate up till then. The questionnaire contained issues such as how the candidate had become familiar with the EUROCERT program, how he/she had used the certificate, whether it had been shown to employers and companies and if yes, whether they accepted it or whether they required some other proof of proficiency in English. The reason for the survey was to

learn about the experiences people had had with respect to the certificate and thereby gaining an impression to what extent the certificate had become accepted up till then.

The second data source consists of two comparisons. Firstly, a comparison was made between two samples, consisting of European candidates, and TSE scores from all over the world. Secondly, an attempt was made to determine to what extent, and in particular with respect to subscores, the EUROCERT sample was comparable to the European non-EUROCERT sample, because of possible variations in averages due to the extent of being prepared for the test.

Scores were gathered only from dates at which EUROCERT administrations take place to ensure that both groups had taken the TSE under identical conditions (EUROCERT and non-EUROCERT candidates are arbitrarily assigned to seats in each test center). For each group, scores from every administration from all of Europe over the years 1988 and 1989, a total of 8 administrations, were mixed. Subsequently, the two samples of rating sheets of 100 examinees each , were manually taken at random from the respective populations. Averages and standard deviations from the samples were compared to averages and standard deviations from TSE score comparison tables covering the periods November 1981 through July 1986 and July 1987 through July 1989.

4 Results

4.1 The survey

From the survey it became clear that the certificate itself had been used in various ways, e.g. to present it to employers and employers-to-be, to show it when applying for MBA institutes, and, of course, colleges and universities in Europe in general. A very important outcome of the survey was that nearly every company or institution, including universities, accepted the certificate as a legitimate proof of an individual's proficiency in English because of the consisting parts. The certificate had been used for application to commercial jobs and apprenticeships in various fields, e.g. research, banking, public relations, tourism, import/export, transportation and automation. Certified examinees confirmed that the EUROCERT certificate, on which scores for the TOEFL, the TWE and the TSE are stated, adds significant value to their C.V.

4.2 Comparison of EUROCERT and non-EUROCERT candidates

Table 7 presents the main data resulting from the comparison between the two samples and the world mean of two different periods. As can be seen from Table 7, the averages of the overall comprehensibility scores from the European samples are consistently higher than the averages from the world comparison tables. This holds true for the subscores on pronunciation, grammar, and fluency in particular. In each of the four

data sets the highest subskill scores were achieved on grammar, followed by fluency and secondly, by pronunciation. The standard deviation for both European samples is smaller than that of the world administrations, for the overall comprehensibility scores and for each of the subskills, thus pointing toward more homogeneity within the European population as can be expected. This assertion is supported by the TSE score manual in which averages on the TSE of 146 countries from all over the world are reported.

Table 7

Mean and (standard deviation) on overall comprehensibility and subskill scores for two samples from administrations in Europe and two periods of world-wide administration.

Administration	N	COMP	PRON	GRAM	FLUE
Europe 1988-1989					
EUROCERT	100	222(36)	2.15(.42)	2.51(.28)	2.21(.42)
REGULAR	100	229(41)	2.20(.42)	2.52(.29)	2.26(.40)
World					
1981-1986	3,500	221(45)	2.10(.49)	2.43(.39)	2.15(.45)
1987-1989	14,365	215(48)	2.11(.51)	2.34(.42)	2.12 (.49)

Table 7 shows that results of the two European samples are more similar to each other than to the results obtained in both averages of world-wide administrations. In fact there are no noteworthy differences, whether the TSE is taken as part of the EUROCERT program or for its regular purpose. Averages on overall comprehensibility and the subskills are practically identical and so are the respective standard deviations.

We carried out a further analysis to investigate the relation of subskill scores at different overall comprehensibility score levels. Figure 2 demonstrates that the regular European TSE candidates with the lowest scores on overall comprehensibility have subscores on the subskill grammar that considerably exceed the other subscores. However, as the score for overall comprehensibility increases, we see that the subscores for grammar, pronunciation and fluency draw closer to each other, yet, the grammar score remains a little higher than the other subscores.

If we compare these results with those from the EUROCERT candidates (Figure 3), it is clear that these observations hold true for both samples. The corresponding lines in the separate graphs follow an almost identical course.

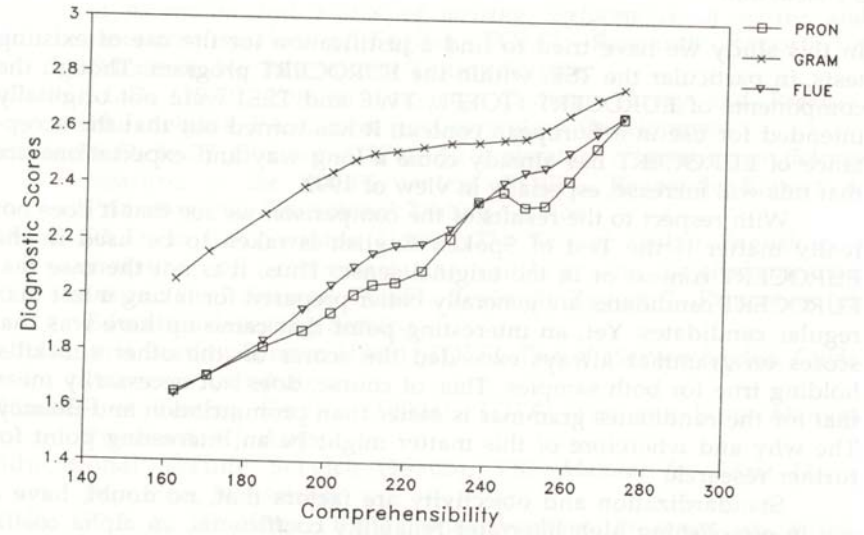


Figure 2
Relation between overall comprehensibility score and subskill scores for Regular candidates

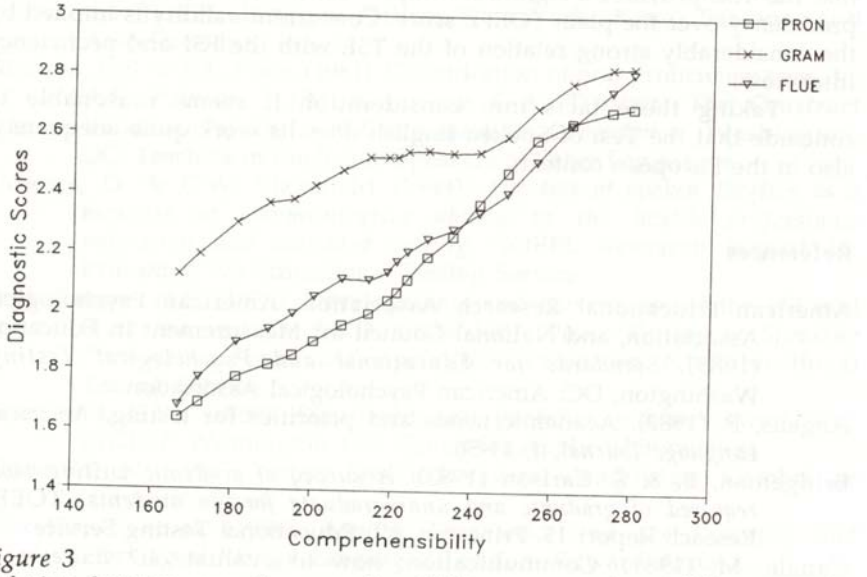


Figure 3
Relation between overall comprehensibility score and subskill scores for EUROCERT candidates

5 Discussion

In this study we have tried to find a justification for the use of existing tests, in particular the TSE, within the EUROCERT program. Though the components of EUROCERT (TOEFL, TWE and TSE) were not originally intended for use in a European context, it has turned out that the acceptance of EUROCERT has already come a long way and expectations are that this will increase, especially in view of 1992.

With respect to the results of the comparison we see that it does not really matter if the Test of Spoken English is taken to be used in the EUROCERT context or in the original sense. Thus, it is not the case that EUROCERT candidates are generally better prepared for taking a test than regular candidates. Yet, an interesting point that came up here was that

scores on grammar always exceeded the scores on the other subskills, holding true for both samples. This, of course, does not necessarily mean that for the candidates grammar is easier than pronunciation and fluency. The why and wherefore of this matter might be an interesting point for further research.

Standardization and objectivity are factors that, no doubt, have a say in establishing high interrater reliability coefficients, an alpha coefficient of .91, and strong intercorrelations among TSE scores. However, a controversial issue was the validity of the test because of its semi-direct character. But as we have seen, also the validity of direct oral tests can be a controversial point.

The moderate correlations with TOEFL scores seem to point out that the TSE provides a significant improvement in the prediction of oral proficiency over the plain TOEFL score. Concurrent validity is implied by the considerably strong relation of the TSE with the FSI oral proficiency interview.

Taking these facts into consideration it seems reasonable to conclude that the Test of Spoken English does its work quite adequately, also in the European context.

References

American Educational Research Association, American Psychological Association, and National Council for Measurement in Education (1985). *Standards for Educational and Psychological Testing.*Washington, DC: American Psychological Association.

Angelis, P. (1982). Academic needs and priorities for testing. *AmericanLanguage Journal,* 1, 41-56.

Bridgeman, B. & S. Carlson (1983). *A survey of academic writing tasks required of graduate and undergraduate foreign students.* TOEFL Research Report 15. Princeton, NJ: Educational Testing Service.

Canale, M. (1981). Communication: how to evaluate it? *Bulletin de l´ACLA,* 77-99.

Carlson, S.B., Bridgeman, B., Camp, R., & J. Waanders (1985). *Relationship of admission test scores to writing performance of native and nonnative speakers of English.* TOEFL Research Report 19. Princeton, NJ: Educational Testing Service.

Clark, J.L.D. (1972). *Foreign Language Testing: Theory and Practice.* Philadelphia, PA: Center for Curriculum Development.

Clark, J.L.D. & S.S. Swinton (1979). *An exploration of speaking proficiency measures in the TOEFL context.* TOEFL Research Report 4. Princeton, NJ: Educational Testing Service.

Clark, J.L.D. & S.S. Swinton (1980). *The Test of Spoken English as a measure of communicative ability in English-medium instructional settings.* TOEFL Research Report 7. Princeton NJ: Educational Testing Service.

Educational Testing Service (1989). *TOEFL Test of Written English Guide.* Princeton, NJ: Author.

Educational Testing Service (1990a). *TOEFL Test and Score Manual.* Princeton, NJ: Author.

Educational Testing Service (1990b). *TSE Manual for Score Users.* Princeton, NJ: Author.

Hale, G. & F. Hinofotis (1981). *New directions in language testing.* Unpublished manuscript. Princeton, NJ: TOEFL Program.

Harris, D.P. (1960). *Testing English as a Second Language.* New York: McGraw-Hill.

Heaton, J.B. (1975). *Writing English Language Tests.* London: Longman.

Kane, H. (1983). *A study of practices and needs associated with intensive English language programs: Report of findings.* Internal report to the TOEFL Program Office. New York: Kane, Parsons, and Associates, Inc.

Madsen, H.S. & R.L. Jones (1981). Classification of oral proficiency tests. In: A.S. Palmer, P.J.M. Groot, & G.A. Trosper, *The Construct Validation of Tests of Communicative Competence.* Washington, DC: Teachers of English to Speakers of Other Languages.

Powers, D. & C.W. Stansfield (1983). *The test of spoken English as a measure of communicative ability in the health professions: validation and standard setting.* TOEFL Research Report 13. Princeton, NJ: Educational Testing Service.

Sollenberger, H.E. (1978). Developments and current use of the FSI oral interview test. In: J.L.D. Clark (ed), *Direct Testing of Speaking Proficiency: Theory and Application.* Princeton, NJ: Educational Testing Service.

Stansfield, CW. (1989). Simulated oral proficiency interviews. *ERIC DIGEST.* Washington, DC: Center for Applied Linguistics.

Underhill, N. (1987). *Testing Spoken Language.* Cambridge: Cambridge University Press.

van Els, T. et al. (1984). Applied Linguistics and the Learning and Teaching of Foreign Languages. London: Edward Arnold.

Wilds, C.P. (1975). The oral interview test. In: R.L. Jones & B. Spolsky, *Testing Language Proficiency.* Arlington, VA: Center for Applied Linguistics.

RESPONSE TO ALEX OLDE KALTER AND PAUL VOSSEN: "EUROCERT: AN INTERNATIONAL STANDARD FOR CERTIFICATION OF LANGUAGE PROFICIENCY

John Read
Victoria University of Wellington

I will comment first on the specific research questions addressed in the paper and then discuss the more general issue of the validity of the EUROCERT programme.

The authors make two comparisons of the performance on the Test of Spoken English (TSE). In this case, European candidates scored consistently higher than did all the candidates on a worldwide basis. This is to be expected, given such factors as the relative closeness of the major European languages to English, the high educational standards in Europe, and the opportunities available to European candidates to use English outside the classroom. If anything, it is surprising that the superiority of the Europeans was not greater.

Secondly, within the European sample, there were "no noteworthy differences" in performance between those who were candidates for the EUROCERT programme and those who were not. It is difficult to see why the non-EUROCERT candidates might have been expected to obtain lower scores. After all, they pay a substantial fee to take the test and are presumably just as motivated to achieve a high standard, which in their case would normally be the cut-off score required for admission to a North American university. Thus the lack of difference seems entirely reasonable.

Beyond these specific questions is the more general issue of whether the EUROCERT programme represents a valid way of using the TOEFL family of tests. The primary purpose of EUROCERT, we are told, is to meet the need for a test that will set internationally accepted and interpretable standards of English proficiency for the purpose of business and trade in Europe. This is rather different from the purpose for which the TOEFL tests were developed and standardized, and yet it seems that they are being used in this new way without modification. No mention is made of any formal attempt to validate the tests for use in the EUROCERT programme. Instead, there is an appeal to face validity, in that major commercial organizations are reported to accept the certificate as evidence that a job applicant has an adequate level of proficiency in English.

It is easy to see how the TOEFL tests would be attractive to personnel managers and other business executives. TOEFL is now well established as an internationally recognized proprietary name; the tests

themselves have excellent psychometric characteristics and are efficiently administered on a worldwide basis to huge numbers of candidates every year. Most attractive of all, each test yields a single score that can easily be interpreted in relation to whatever cut-off point the test user has set. In the case of EUROCERT, though, we are not given any information about how the cut-off scores were determined. It may indeed be true that those candidates who qualify for certification, and especially those "Certified with Honors", are highly proficient and would do well in any English proficiency test. However, when candidates fail to achieve the minimum scores, it is an open question whether they really lack proficiency or are being penalized by the use of an inappropriate measure of their competence in the language. In the absence of any kind of validity data, the meaning of Eurocert results is difficult to evaluate.

If an international certification programme is so important for European business corporations, it would make more sense from a language testing viewpoint to develop a new test for the purpose, or at least to produce a modified version of TOEFL that would reflect the requirements of European employers and the uses of English in European business and industry. One current trend in language testing is the more serious attention being paid to content validity, especially in specificpurpose tests. The TOEFL programme itself provides a good example of this trend, in the care that was taken to identify and select suitable writing tasks for the Test of Written English (TWE), based on a large-scale survey of academic staff in North American universities. Thus, it is rather ironic that the TWE is now being used in a European programme for which the tasks may be quite inappropriate. The appropriateness of the tasks is a matter for empirical investigation; it should not simply be assumed.

EUROCERT represents one way of meeting the apparent need in Europe for an internationally recognized certificate of English proficiency, but it is a disappointingly limited one. The TOEFL tests are conservative in various respects, in comparison to the more innovative approaches to the testing of English for specific purposes that are found in recent British tests, for example. Therefore, EUROCERT in its current form can be seen as a lost opportunity to develop a state-of-the-art test for the purpose. At the very least, the use of the TOEFL tests in this way should be properly validated.