DOCUMENT RESUME

ED 326 052 FL 018 970

AUTHOR Laurier, Michel; Des Brisay, Margaret
TITLE An Integrated Approach to Developing Small-Scale Standardized Tests.
PUB DATE Apr 90
NOTE 19p.; Paper presented at the Meeting of the World Congress of Applied Linguistics sponsored by the International Association of Applied Linguistics (9th, Thessaloniki, Greece, April 15-21, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computer Assisted Testing; Computer Oriented Programs; Computer Software; English (Second Language); English for Academic Purposes; Foreign Countries; French; *Language Tests; Second Language Learning; *Standardized Tests; Student Placement; Teacher Developed Materials; *Test Construction; Test Validity
IDENTIFIERS *Canadian Test of English for Scholars and Trainees; Summer Language Bursary Program Placement Test

ABSTRACT

A discussion of the use of standardized tests in second language assessment focuses on the usefulness of program-based language tests aligned with specific curricula and capable of motivating students to participate in classroom activities that are a preparation both for the test and for life after the test. The purpose of this paper is to show that a small-scale test development project can be successfully developed using relatively limited resources, inexpensive computer software, and expertise that can be reasonably acquired by second language teachers. Experiences in developing, pilcting, and validating two small-scale standardized tests, the Canadian Test of English for Scholars and Trainees and the Summer Language Bursary Program Placement Test are used as illustrations. Standardized tests are defined, and the process of developing standardized measures is outlined. Applications of classical test theory and item response theory in this context are explained, including the use of sample-free calibration, test-free person measurement, goodness-of-fit measures, and multiple reliability. The role of informed judgment as an adjunct to statistical analyses in test development is also discussed. A brief bibliography is included. (MSE)

# An Integrated Approach to Developing Small-Scale Standardized Tests

Paper to be presented at the 9th World Congress of Applied Lingusitics, Thessalonika, Greece, April 14, 1990

Michel Laurier
Department of French
Carleton University
Ottawa, K1S 5BG
CANADA

Margaret Des Brisay
Second Language Institute
University of Ottawa
Ottawa, K1N 6N5
CANADA

---

One of the most vexing problems facing language program planners stems from the fact that assessment activities used in the classroom are very different from the procedures used on what is commonly referred to as "a standardized test." Of course, we should not expect classroom evaluation to be identical with standardized procedures (Oller 1979). A good classroom test should always be designed for a specific group and should have immediate relevance for the teaching content. However, pedagogically appropriate assessment procedures are not usually reliable enough to satisfy the information requirements of interested parties external to the classroom context. The usual solution is to employ a commercially available standardized test to provide the information necessary for admissions officers, receiving institutions, funding agencies and program evaluators. This risks creating a further problem in that inappropriate and unfamiliar instruments may be used to measure achievement of program objectives.

One solution to this dilemma is for programs themselves to develop standardized instruments that are aligned with their curricula and which will motivate the students to participate in classroom activities that are a preparation both for the test and for life after the test. The recent "proficiency movement" headed by the ACTFL, claims that there is no reason not to teach for the test if the test represents a situation in which the learner can demonstrate what he/she can do with the language: "We test what we teach, we also teach what we test" (Magnan 1985).

It is the purpose of this paper to show that a small-scale test development project can be successfully undertaken using relatively modest resources, inexpensive software packages and expertise that can be reasonably acquired by ESL/FLS staff members. The paper is based on the experience of the two authors in the development, trialling and validating of two small-scale standardized tests: the Canadian Test of English for Scholars and Trainees (CanTEST) used by the Canadian International Development

Agency to select and certify candidates for academic or professional exchanges in Canada and the Summer Language Bursary Program (SLBP) Placement Test which will be offered to participating colleges or universities to help them assign young Canadian anglophones to various levels of French instruction in special summer programs funded by the Secretary of State. The CanTEST and the SLBP Test can be considered as small-scale tests. They are not likely to be administered to more than 1200 examinees per year - an average of 100 per month. The human resources to develop and maintain these tests represent one person-year plus part-time clerical support.

The SLBP Test is still under development. Since it will be used as a pretest and a posttest as well as for placement, at least two parallel versions are required. In line with program objectives, only speaking and listening skills are to be measured. The first two parts can be seen as a routing test towards a confirmatory oral interview. The content of the whole test will reflect the communicative approach currently used in the program - or that the program co-ordinators wish to have used. Since there are very heavy practical constraints on the administration and scoring of the test, a multiple choice format has been used. The test project was initiated because none of the standardized instruments that were available in French could meet all of these requirements. Both Classical Test Theory (CTT) and Item Response Theory (IRT) have been used since the early stages of development as a way of improving reliability and of equating the two versions.

The CanTEST, on the other hand, has been used operationally for over five years in the People's Republic of China to certify candidates for academic and professional exchanges with Canada. The listening and reading sub-tests of which the CanTEST item bank is composed have been analyzed following both field-testing and operational use using CTT. The computer program used for the analysis, ITEMAN (Assessment System Corp. 1987), is inexpensive and extremely easy to use. All the traditional indices for CTT are provided permitting the developers to identify items with poor psychometric properties as defined by CTT. The population for the CanTEST has been extremely stable with respect to the range of abilities represented. Questions of cultural bias did not arise since all the testees were of very similar linguistic, cultural and educational backgrounds. The possibility now exists, however, that the item bank will be used to generate tests for a more diverse population and developers feel the need to confirm the information provided by CTT through the application of IRT.

It is important to begin the report of our experiences by defining the term "standardized" and specifying the testing situations which justify the effort and expense of standardizing. According to Millman & Greene (1989:340), "a standardized test is one for which the conditions of administration and scoring procedure are designed to be the same in all uses of the test."

In addition, norms or interpretive guides must be available to give meaning to the scores so that there can be comparability of results from one test administration to another regardless of the versions used or the proficiency of the particular group being tested. There is clearly no need to standardize in the case of teacher made tests since the necessity for comparability does not extend beyond the classroom. However, in the case of the two tests under discussion there was such a need. The CanTEST is administered to different groups of candidates whose future will be affected by consequential decisions based on their score. The SLBP Test will allow students to transfer course credits from one institution to the other and will assess overall program effectiveness.

As any other test, a standardized instrument must be valid, reliable and practical. The validity requirement is met when the test actually measures what we intended to measure. Reliability refers to the extent to which a test will yield similar results on the same subjects over repeated administrations. In order to compare results as we do, it is very important that our instruments be reliable. Practicality is related to the resources that are needed for the administration and the marking of the test. Concerns about reliability and practicality have generally led to the wide spread use of multiple choice items on most standardized tests. The validity of this item format has been challenged (Shohamy & Reves 1985). Multiple-choice answering is not an authentic language task and the format can easily be seen as a way to reduce language proficiency to a set of discrete items instead of viewing it as a process in which various competencies must be integrated. However, as far as receptive skills are concerned, well-constructed multiple choice items certainly measure how well a testee can check the various hypotheses as represented by the different options. Oller (1978) has stressed the importance of hypothesis confirmation processes in second language proficiency.

The steps involved in the construction of a standardized test are well described in Tinkerman (1971:46):

- Developing the test specifications: First, we must specify the purpose of the test, the content and the item format. At this stage, careful attention must be paid to the planning of the test construction procedure;

- Writing the test items: To avoid discarding too many items, the first version of the test must be written carefully and thoroughly reviewed. There is an abundance of literature on the creation of good items. For many language tests, recording sessions may needed.

- Pre-testing the items and analyzing the item statistics: This pre-experimentation can be conducted with few subjects (100 is usually sufficient). A first screening can be done using standard item analysis procedure. Defective items can

be either modified or rejected. The test should also be compared with other data that may be available.

- Compiling the preliminary test forms.

- Trying out the preliminary test form for standardization purposes: Depending on the purpose of the test, the nature of the items and the psychometric model, the sample size could range from 200 to 2000 examinees; administration procedures must be carefully controlled and various statistical analyses will be conducted to ensure reliability and validity.

- Preparing norms, a test manual and supplementary materials: If there are different versions they must be equated; the administration procedure must be very clear.

- Printing and publication: This include the duplication of the test material (written and oral) and the distribution of the instrument.

Classical Test Theory

CTT was used by both developers in the early stages to provide information for the refining and possibly rejecting of test items and in fact, as noted above, the CanTEST item bank was compiled using informed judgement and CTT alone.

The theory is based on the fact that any type of measurement is never absolutely error-free. Therefore, a score on a test does not necessarily represent the subject's "true score". Thus the basic equation is:

True score = Observed score + Error

The problem is that the amount or error cannot be predicted; the theory can simply determine a range of values that the error component is most likely to take.

Underlying this basic equation are two important assumptions. First, the theory works with a score, that is to say, with the proportion of correct answers on a given number of items measuring the construct. The examinee's ability is always expressed as score which score can be converted later on different scales (percentage, Z-score, percentile...).

Secondly, the theory assumes that the error variance is constant throughout the entire ability range. Any test developer is aware of the fact that it is easier to create items that discriminate at the intermediate level than items that discriminate at the advanced or beginning levels. Yet, the theory implies that the range of error will be the same at any level. Reliability indices (Cronbach's alpha, KR-20) that are usually reported for a test are very useful as an overall indication of the test

efficiency but they may disguise the fact that the test lacks discriminatory power at advanced and beginning levels of proficiency. We will return to this point later.

The basic equation of the classical analysis also means that the concerns of CTT have focussed primarily on the reliability of a test rather than with its validity. Many correlational techniques have been developed within the classical theory framework but most of them are of limited use for language assessment because the criterion can be always questioned. All we can do is to see how well a test correlates with an acceptable external measure but even this begs the question of whether either measure was a valid one of language proficiency.

Test developers have also been constrained by the fact that all the indices provided by CTT depend on the sample used. To make sure that all the different levels of proficiency are represented and that test scores will not be affected by any cultural, sociological or linguistic bias, the field-testing must be conducted carefully. The sample must always reflect the population with which the test will be used to make decisions. If the sample is adequate, the proportion of correct answers (p-value) and the biserial are good indicators of the difficulty and discrimination of a given item. However, the fact that these indices are sample dependant may represent a serious limitation for applications such as item banking or test equating. The recommended equating procedures (same two tests, same population) are rarely practical.

Here is an example of the use of difficulty and discrimination index for item analysis. As a part of the SLBP Test in French, we wanted to construct a sub-test to measure how well a learner was able to choose the statement that was the most appropriate for a given situation.

For example, the subject may be presented with this situation:

> You are in a train. You do not know the passenger who is sitting beside you and you wonder if you may smoke. The person is a man, about 50 years old; he is reading a magazine. To inquire which question would you use?
>
> A- Est-ce que tu veux que je fume?
> B- Auriez-vous l'obligeance de me permettre de fumer?
> C- Est-ce que cela vous dérange si je fume?
> D- Vous permetteriez que je fume?

The preliminary test comprised 50 of these items. We wanted to keep 30 of them in two different versions of 15 items each. Various computer programs are available to calculate discrimination and probability indices: LERTAP, Microcat/ITEMAN, TESTAT, SPSS/Reliability...

Figure 1 shows some items that were rejected because of their poor discrimination index or because they were clearly too difficult or too easy.

| Item number | 7 | 21 | 28 |
|---|---|---|---|
| Correlations | | | |
| - Point-biserial | .276 | .221 | .174 |
| - Biserial | .514 | .294 | .224 |
| Probability | .925 | .716 | .359 |
| Number of answers | | | |
| - option a | 1664* | 36 | 401 |
| - option b | 59 | 1288* | 646* |
| - option c | 17 | 357 | 158 |
| - option d | 47 | 88 | 582 |
| - omitted | 10 | 28 | 10 |

Figure 1: Bad items

Point-biserial and biserial correlations indicate the relation for every student, between the result on a particular item (right or wrong) and the total score. These values ranges from -1 to 1. A low index means that many advanced students who should have selected the right option did not and beginners who should have failed the item got it right. On the contrary, a high index means that the item discriminates well between the advanced students and the beginners. Item 21 and 28 were rejected because of their poor discrimination. Point-biserial and biserial give the same information but biserial correlations are less affected by extreme values of the probability index. The probability index represents the proportion of examinees who got the item right. With more than 9 out of 10 examinees choosing the right answer (a), Item 7 was eliminated from the final version as being too easy. The probability index is calculated from the total number of answers given to the different option. below. These figures may help to expose bad distractors. For example, we observed that the option d on item 28 was too attractive in relation to the right answer "b".

| Item number | 9 | 17 | 18 |
|---|---|---|---|
| Correlations | | | |
| - Point-biserial | .363 | .503 | .632 |
| - Biserial | .605 | .631 | .813 |
| Probability | .890 | .478 | .649 |
| Number of answers | | | |
| - option a | 25 | 859* | 172 |
| - option b | 6 | 208 | 253 |
| - option c | 159 | 53 | 145 |
| - option d | 1601* | 666 | 1167* |
| - omitted | 6 | 11 | 60 |

Figure 2: Good items

On the other hand some items were particulary effective because they had high discrimination index . As can be seen from Figure 2 (page 6), Item 9 is easy and item 17 is difficult. They both discriminate well even though they have a distractor that is less effective (b in item 9, c in item 17). Item 18 is a good example of an ideal item at the intermediate level.

We obtained a reliability index of .867 (Cronbach's alpha) on the preliminary test. This is an acceptable value but we needed 50 items to reach this level. The theory predicted that with 15 items the reliability coefficient would drop to .662, which means that the error component would be too large. However, by selecting the most efficient items, we were able to raise it to .781 (version A) and .751 (version B).

## The Item Response Theory

IRT has been described as "undoubtedly the most striking development of the past several decades in educational measurement" (Carroll, 1990). The theory assumes that there is an underlying trait on a test being measured by the set of items and that the accuracy of the measurement at a given level of ability depends on certain characteristics of the items, the item parameters. In other words, IRT considers the item, not the score (proportion correct).

Various models have been proposed using one, two or three parameters. The best known one is the one parameter model which is also known as the Rasch model. Although it is the least accurate for multiple choice applications, it is the most interesting for small scale language testing. First, it is the most accessible one. A Rasch analysis can be conducted with small-samples (200 subjects min.) whereas a three-parameter calibration would require at least 1000 examinees. Second, under the Rasch model, the number of right answers on a common test represents the best estimation of a subject's ability. Since in most cases number-right scoring is the only practical way to mark a test, this property makes the Rasch model very convenient. Third, more

sophisticated models are very complex. The one-parameter model is the most mathematically tractable. It is based on a simple relationship between the probability of a right answer and the subjects' ability. The calibration process consists in finding the ability level where the item is the most informative. This value corresponds to a single parameter, the item difficulty.

The information obtained from IRT is a useful alternative or complement to CTT from various points of view which are discussed below.

Sample-free item calibration: Since the calibration procedure aims at fitting a curve rather than simply calculating proportions of correct answers, the difficulty index corresponds to the value that an item is most likely to take for a pattern of answers. Therefore, the difficulty indices are less affected by imperfect distribution of the population. Although the sample should represent the target group, sample-free calibration makes the field-testing a lot easier.

With sample-free calibrations, a large item bank can be created by trialling various sets of items with different samples of the population.

Figures 3 and 4 show an application of the Rasch analysis for a sub-test of listening comprehension in French. We constructed two tests with four passages, each followed by 15 comprehension questions. Because the Rasch model uses generally a "logit" scale (from -4 to 4), the mean of the difficulty indices of the 15 questions can be calculated and compared with the mean of the other set of questions. On Version A the question difficulty for the first and the last passages were the same. We decided to eliminate the last one. On Version B we dropped the last passage which gave us questions that were too difficult. We also deleted questions on the third passage that were too easy.

It should be mentionned that the term sample-free calibration may be somewhat misleading. The scale is undetermined so that one cannot compare difficulty levels from Version A with those from Version B. In order to do so, we should have included some common items, called "linking items", to anchor the scales on a common zero point.

**Figure Three: Listening Comprehension - Version A -Summer Language Bursary Program, Part I.**

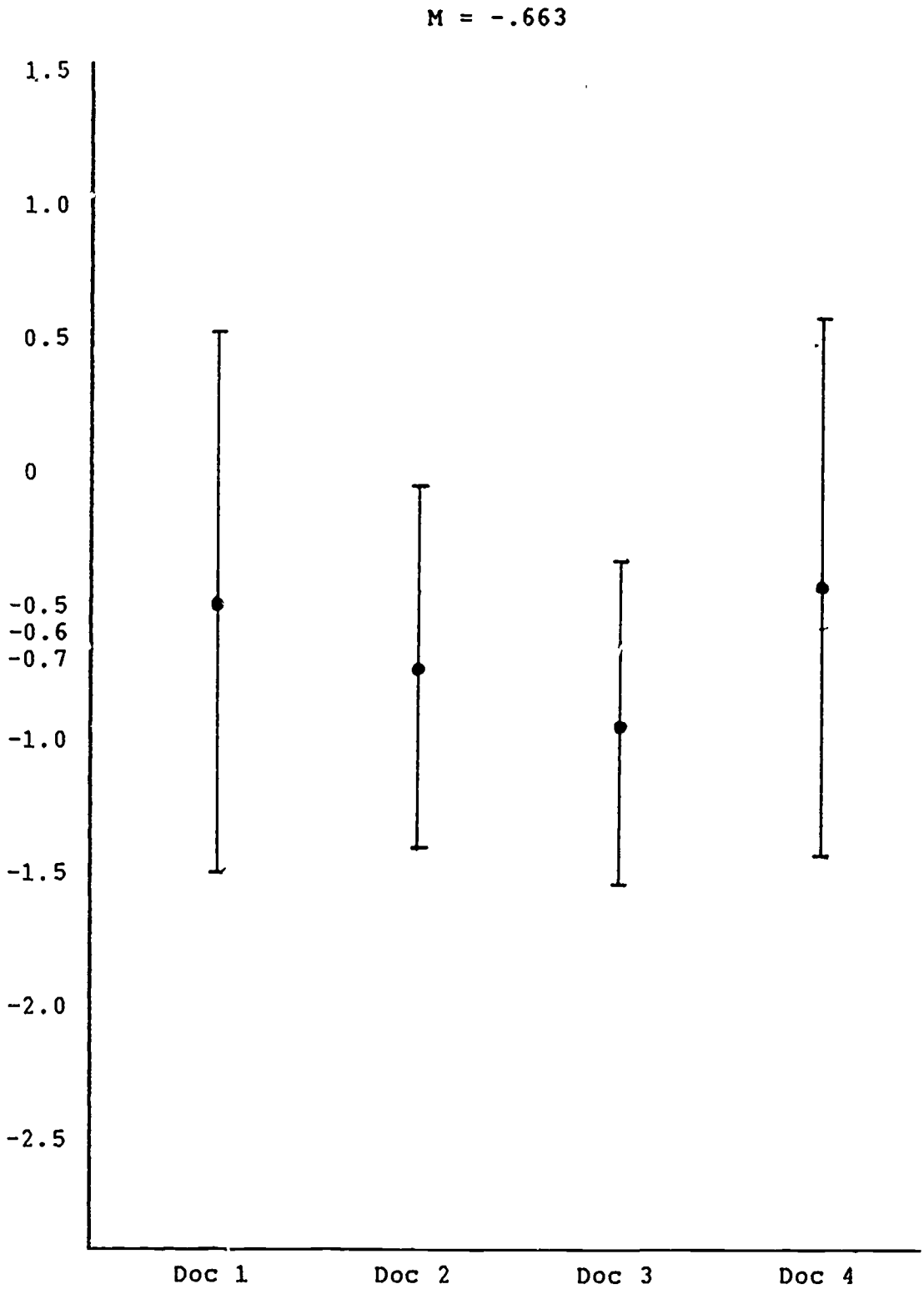
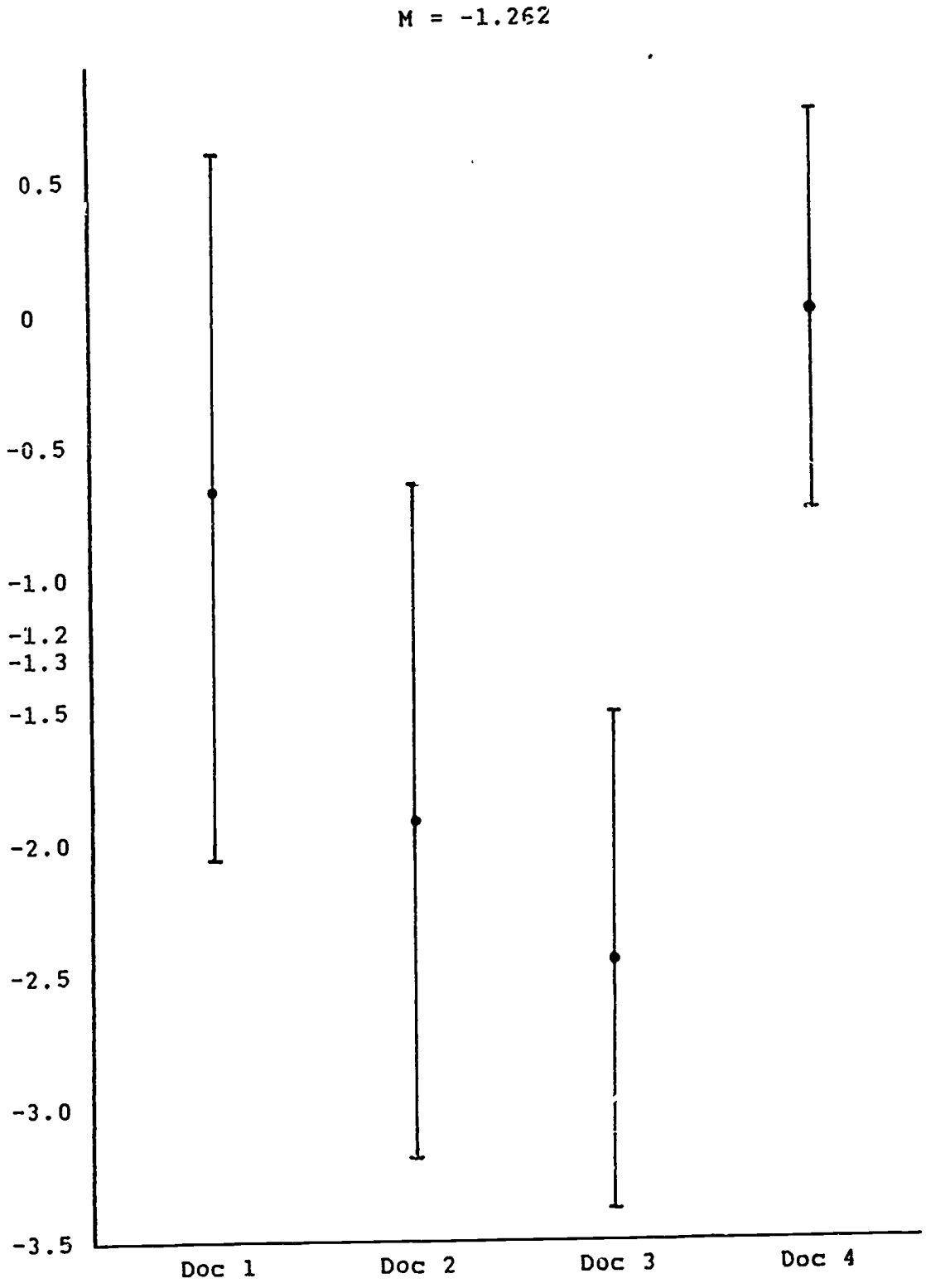


**Figure 4: Listening comprehension - Version B - Summer Language Bursary Program, Part I**



Test-free person measurement: Once the items have been calibrated, we can estimate an examinee's ability from his/her answers to different sets of items. This property is particularly interesting for adaptive testing. With adaptive testing, different subjects may be presented different items according to their own level but the final estimations obtained with the appropriate mathematical estimation procedures can be compared. Test-free person measurement also permits us to design various versions of a test using different items drawn from a item pool and to equate them without any conversion tables.

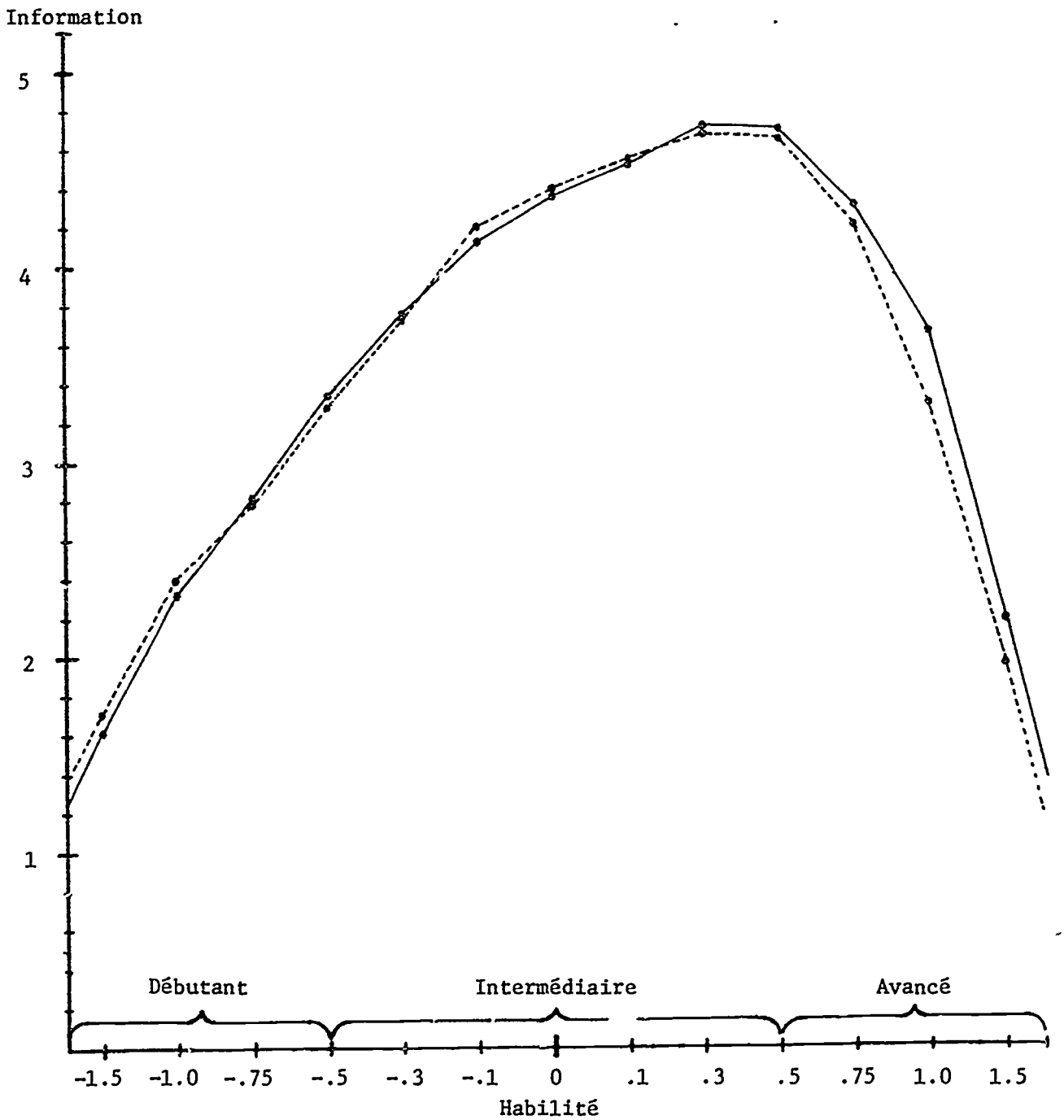
Measures of goodness-of-fit: Since the calibration procedure tries to fit a curve, some algebraic formulae have been developed to measure how well the data fits with the curve. Misfitting items are often bad items which will have to be rejected. At the subject level, finding deviant subjects (or examinees) whose answer patterns do not fit with the model may be a way to detect cheaters, people with different linguistic or cultural backgrounds, and other special cases.

Multiple reliability: The theory provides an information function of the parameters that indicate how efficient an item is at a given point on the ability scale. The basic idea is that an item that is too difficult or too easy with respect to the examinee's ability is not relevant. When no guessing takes place, the maximum information is obtained when there in a 50% chance of a subject getting the right answer.

The theory also says that the information is additive. In other words, we can sum up the information obtained with the items of a test at different levels and plot an information curve. As the information increases, the error decreases. The information function is a useful alternative to the classical reliability because it tells us at which point the test is most reliable while indices such a Cronbach's alpha, as noted above, assume that error is equal at all levels. Thus we can design tests that will be the most informative at the point where the crucial decisions will be made.

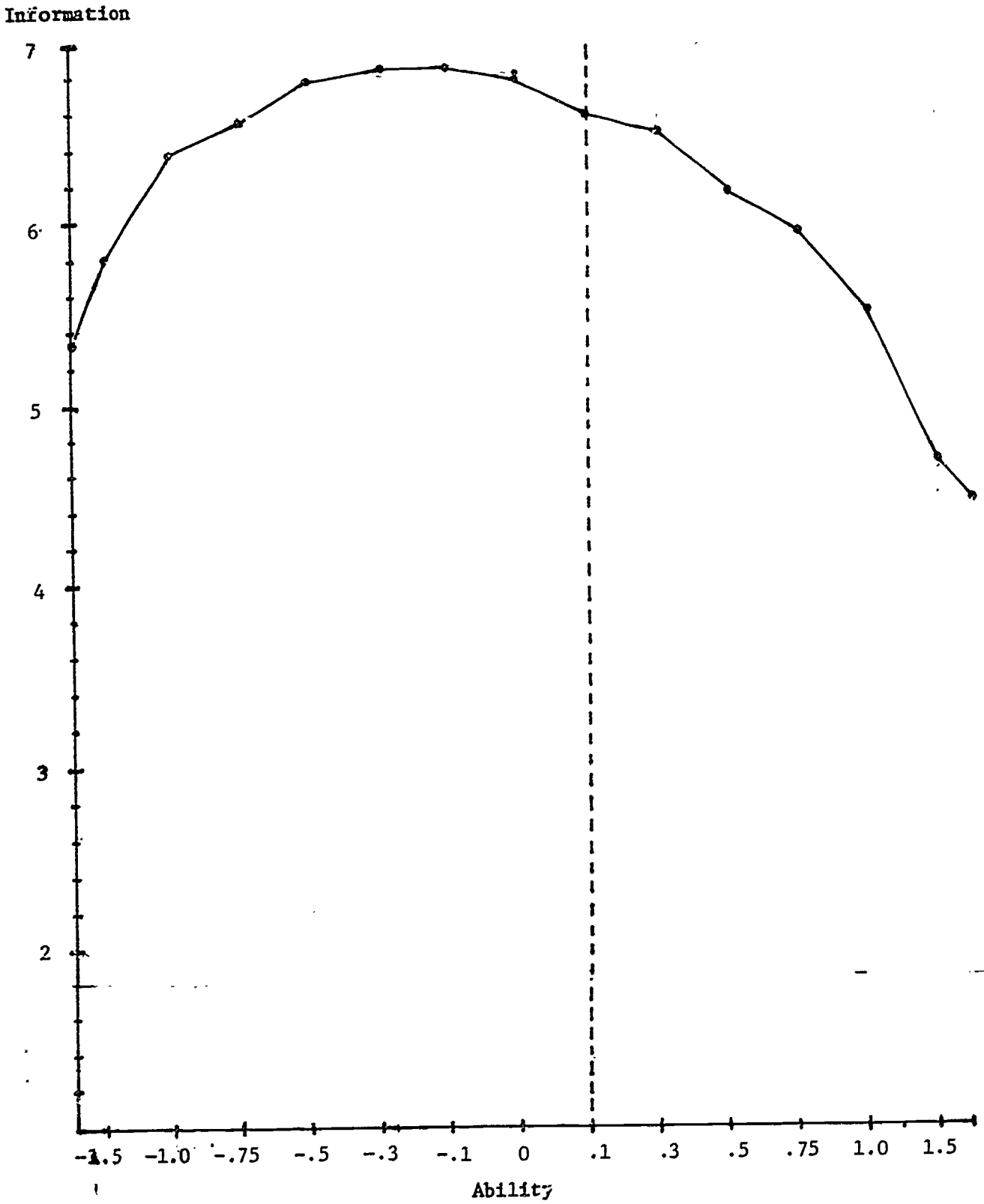
IRT proponents also claim that two identical information curves mean that the tests could be considered parallel. Figure 5 on page 12 shows the information curve of the two 15-item tests that were created from the experimental instrument in which the student were ask to find the most appropriate statement. Combining different items, we manage to get two similar curves in order to get two equivalent versions. On both versions of the test, the maximum information is obtained at the "high intermediate level".

Figure 5: Information curves for two sets of 15 selected items
SLBP Test - Part II




Version A: ————— Version B: ----------

Figure 6:: Information curve for CanTEST Listening (50 items)



In the case of the CanTEST, the information curve was used to confirm how well the test is performing rather than to equate two versions of the test. The CanTEST is primarily a selection test that aims at certifying candidates who are ready to participate in a Canadian attachment (academic or professional) or who could be brought to such a level within a 15 week intensive training period. The test does not have to be very effective at the extremes of the ability scale but we want to make sure that candidates without the potential to succeed are not admitted to the program. Figure 6 on page 13 shows the information curve of the Listening Comprehension test. As a selection test, the CanTEST is a peaked test. The peak of the curve is slightly below the cut-off point where the consequential decisions are being made. Since we do not need any information at the very advanced level, the information curve drops rapidly after the cut off point.

Informed judgement

Informed judgement plays a role in every phase of test development: in selecting and designing test content, then in selecting the most appropriate psychometric approach and in interpreting the statistical information obtained but most importantly in determining the inferences that can be made on the basis of test scores, what Messick (1981) calls "the meaning of the measure."

The content of a test can only be specified by people who made an effort to define what language is and how it is learnt or acquired. A language test must correspond to our conception of the language and the methodologies that are currently used in the language class. This is why "informed judgements" from people involved in language education is essential. Models such as those proposed by Canale and Swain (1980) or Bachman (forthcoming) must be fully validated in order to establish the language testing on theoretical foundations. These models must not only match with the data obtained but they must also reflect teachers' and students' intuitions.

Here is an example of what can happen if statistical properties alone are used to determine the quality of an item. In a preliminary version of the French placement test, we had the following item:

> Your want to see your French teacher. There is a note on the door saying that she will be back at her office at 3:00. It is now 2:45. What will be on the door.
>
> A) Le professeur sera de retour dans trois heures.
> B) Le professeur est revenu il y a quinze minutes.
> C) Le professeur est revenu depuis trois heures.
> D) Le professeur sera de retour dans quinze minutes.

The last option (The teacher will be back in fifteen minutes) is the right answer but a student would not know how long to wait if such a message was posted on his/her teacher's door! Even if this answer makes no sense, it was an exceptional item from a statistical point of view. With a probability level of .36, it showed a biserial correlation of .63 and the fit with a three-parameter model was excellent. However it clear that on the "common sense" basis, this item was a bad one.

The IRT analysis of the CanTEST Listening Comprehension test provides additional examples of the role played by informed judgement in the application of statistical information. The analysis of the listening test identified a few misfitting items.

Item 7 (on the overhead) has virtually identical properties to Item 6 using CTT but only Item 6 fits the IRT model. The Item Characteristic Curve (ICC) suggests a guessing effect as some low-level testees appear to have gotten this item correct. An examination of the answer sheet shows that the correct answer for Item 7 was "a". Since there had not been an "a" since Item 1, it is quite probable that the guessers had successfully applied a test taking strategy common to multiple choice tests. Had the correct answer been "b", or had the item appeared elsewhere in the test, the figures could well have been different so misfitting or not, there seems no reason to discard this item. One is reminded however, once again, of how the position of an item in the test affects its psychometric properties.

Item 28 is a write-in answer, the only one for this sub-test which required testees to record a specific number (the length of a crate). The ICC suggests that low intermediate students did better than high intermediate students on this item although the item as a whole was very easy one. But as was mentioned above, the CanTEST is aligned with an instructional program and picking out specific facts and figures from a spoken text are stressed in the program. Such questions must be on the test to motivate students to attend to classroom activities designed to develop this skill.

Two other items that were considered misfitting were in fact, excellent items that were over-fitting the model. One of them, Item 48 is shown on Figure 7. If we refer to the discrimination indices provided by CTT, we realize that these two items discriminate very well (biserials of .79 for Item 48. However, since the Rasch model assumes equal discrimination for all items, very effective items will be labeled as misfitting. In that case, CTT or informed judgment will both warn us not to discard these good items. It should be further noted that this would not happen with a 3 parameter model which assigns a discrimination index to the items. Unfortunately, the 3 parameter model is not always a

practical solution to the small scale test developer because of the large sample size which is required.

As we have already mentioned, the experience of the CanTEST developers contrasts with that of the developers of the SLBP Placement Test in that IRT was not used in the test construction phase and is presently being used mainly to confirm the effective functioning of the test. Nor is it likely that IRT will ever play as influential a role in the selection of test content for the CanTEST as for the SLBP Test since considerations of washback on the instructional program for which the CanTEST serves as the exit test and of the cross-cultural context in which testing takes place must continue to be balanced against considerations of maximum efficiency.

All IRT models have serious limitations. Any application of the theory assumes that there is a latent trait, i.e. a major component that is common to all the items. This property is known as the unidimensionality of the test. To a certain extent CTT also assumes the unidimensionality; under IRT, this requirement is essential. For language testing, this issue is fundamental: Is language proficiency unidimensional? The unidimensionality of a language test is sometimes difficult to establish. Even though statistical procedures are available there will always be a need for some kind of "informed judgement". Fortunately, recent studies have also prove that IRT is more robust than what we had thought (Harrison 1986) and that many language test show a dominant dimension (Henning Hudson & Turner 1985). However this principle is a major constraint on the format of a language test. One aspect of unidimensionality is the local independence of the items. Local independence involves that a correct answer should not provide any clue for another item. This principle is hardly compatible with the idea of building hypotheses using information from the context. For this reason IRT is probably not suitable for a Cloze test.

One may ask whether it is realistic to believe that language performance can be analyzed with a model that is so restrictive. Traub (1983) wonders if we should narrow our conception of educational measurement to meet the unidimensionality requirement. This problem of over-simplification is particularly obvious with the Rasch model. Informed judgment warns us that the items on a test never discriminate equally and that on a multiple choice test there is also some guessing taking place.

What the language tester must realize is that statistical analysis and psychometric theories are tools that help us to construct good tests but they should never dictate what should be tested. However, since a good test must measure what it is supposed to measure a detailed content description, an interesting format or a wide variety of tasks does not necessarily ensure that the test will be acceptable. In certain situations, reliability is a critical consideration. We believe that it is possible for

small-scale test developer to obtain valid and reliable tests by integrating three approaches: CTT, IRT and "Informed judgment". In our experience all of them are required to give a full picture of how effective a test is. The tools that we have described in this paper help the test designer to minimize the error of the instrument. And, of course, error is not what we want to measure!

BIBLIOGRAPHY

Assessment System Corp. (1987). User's Manual for the Microcat Testing System, 2nd ed. St. Paul, MN.

BACHMAN Lyle (forthcoming). Fundamental Considerations in Language Testing. Reading, MA: Oxford University Press.

CANALE Michael & Merrill SWAIN (1980). "Theoretical bases of communicative approaches to second language teaching and testing". Applied Linguistics, 1, pp 1-47.

HARRISON David A. (1986) "Robustness of IRT parameters estimation to violation of the unidimensionality assumption". Journal of Educational Statistics, 11(2), pp 91-115.

HENNING G., HUDSON T. & J. TURNER (1985) "Item Response Theory and the assumption of unidimensionality for language tests". Language Testing, 2(2), pp 141-154.

CARROLL John B. (1990) "Future development in educational measurement". In Walberg H.J. & G.D. Haertel (eds) The International Encyclopedia of Educational Evaluation, New York: Pergamon Press, pp 245-250.

MAGNAN Sally Sieloff (1985) "From achievement towards proficiency though multi-sequence evaluation". In James C.J. (ed) Foreign Language Proficiency in the Classroom and Beyond. Lincolnwood, IL: National Textbook (with ACTFL) pp 117-146.

MESSICK Samuel (1981) "Evidence and ethics in the evaluation of tests". Educational Researcher, november, pp 245-250.

MILLMAN Jason & Jennifer GREENE (1989) "The specification and development of tests of achievement and ability" In Linn R.L. (ed) Educational Measurement, 3rd ed. New York: American Council on Education and Macmillan Publ. Co., pp 330-350.

OLLER John W.JR. (1978) "Pragmatics and language testing". In Spolsky B. Advances in Language Testing, Serie 2: Approaches to Language Testing. Arlington, VA: Centre for Applied Linguistics, pp 39-57.

OLLER John W.Jr. (1979) Language Tests at School. London: Longman.

SHOHAMY Elana & Thea REVES (1985) "Authentic language test: Where from and where to?". Language Testing, 2(1), pp 48-59.

TINKERMAN Sherman N. (1971) "Planning the objective test". In Thorndike R.L. (ed) Educational Measurement, 2nd ed. Washington, DC: American Council on Education, pp 46-80.

TRAUB Ross E. (1983) "A priori considerations in choosing an item response model". In Hambleton R. (ed) Application of Item Response Theory. Vancouver, BC: Educational Research Institute of British Columbia, pp 57-70.