DOCUMENT RESUME

ED 362 006 FL 021 386

AUTHOR Blais, Jean-Guy; Laurier, Michel
TITLE The Dimensionality of a Placement Test Components.
PUB DATE Sep 93
NOTE 28p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Foreign Countries; *French; Higher Education; Language Proficiency; *Language Tests; Reading Comprehension; Second Language Instruction; *Second Languages; Test Format; Testing
IDENTIFIERS Canada; *Placement Tests

ABSTRACT

A computerized adaptive test for placement of students in postsecondary French second language courses is evaluated for unidimensionality of its three component tests: reading comprehension of a short paragraph; selection of the appropriate statement in a given situation; and a "fill-in-the-blank" section. A variety of statistical procedures were used to assess the components' unidimensionality, including a structural equation approach, factor analysis, nonparametric approach, and item response theory approach. Results of each of these analyses are explained and synthesized. It is concluded that the varying and sometimes conflicting results raise questions about the concept of unidimensionality, and that unidimensionality is a matter of degree rather than a yes/no issue, depending heavily on expert judgment. A 49-item bibliography is included, and procedures and results of the different analyses are appended. (MSE)

The Dimensionality of a Placement Test Components

Jean-Guy BLAIS
University of Montreal
Faculty of Education
Dept of Studies and
Management in Education
P.O. Box 6128, Station A
Montreal (Quebec)
Canada H3C 3J7
Tel: 1(514)343-7527
Fax: 1(514)343-2497

Michel LAURIER
University of Montreal
Faculty of Education
Dept of Studies and
Management in Education
P.O. Box 6128, Station A
Montreal (Quebec)
Canada H3C 3J7
Tel: 1(514)343-7034
Fax: 1(514)343-2497

Paper submitted for publication in ***Language Testing***

September 1993

The Dimensionality of a Placement Test Components

Jean-Guy BLAIS and Michel D. LAURIER, *Université de Montréal*

In order to find the most appropriate group for students enrolling FSL classes at the post-secondary level, a placement test has been designed. The purpose of this test is to measure general proficiency in the language. Using IRT procedures, an adaptive version has been developed. The test consists of three parts: comprehension of a short paragraph, selection of the appropriate statement in a situation, and a "Fill-the-gap" part. Different approaches to assess the unidimensionality are used: structural equation approach (LISREL), factor analysis approach (TESTFACT), nonparametric approach (Stout's procedure) and IRT-based approach (Bejar's procedure). These approaches are complementary. The subtests generally meet the unidimensionality assumption for IRT although they measure more than one ability.

KEYWORDS: **Unidimensionality, placement test, adaptive testing**

Introduction

Because of the particular linguistic situation in Canada, many colleges and universities are faced with the problem of assigning the most appropriate group to students enrolling in a French as a second language course. Some students have a very basic knowledge of the language either because they drop French at high school or they did not study in Canada. Some others are very fluent; they may be students who learnt the language in French-speaking communities, or immersion students or talented core French students.

In order to solve this placement problem, a computer adaptive test has been set up (Laurier, 1991). Multiple-choice items have been calibrated, from a paper-and-pencil version, using an item response theory model (3-parameter logistic model) and were stored in three small item banks. The decision to create three different banks was based on the nature of the tasks rather than on a full dimensionality study. Each bank is used for a subtest and the final result is the combination of the ability levels on the three subtests.

The purpose of this study[1] is to determine if these three subtests are unidimensional. The study will be used as evidence for the construct validity of the test and to determine whether IRT procedures can be reasonably applied. Then, it will be possible to plan the addition of new items in order to expand the item banks so that the whole range of examinees' ability could be adequately covered. It will also be possible to improve the way the three subtests results are aggregated to facilitate the interpretation of this compound result for an appropriate placement decision. Finally, this study is a prerequisite for the development of additional subtests in the CAT version. For example, one may claim that the inclusion of a listening part would probably measure an important trait that is not measured in the present test.

---

[1] This study has been funded by the Social Sciences and Humanities Research Council of Canada (#410-92-1400)

In addition to the theoretical aspects of language test structure, there are two practical reasons for using a unidimensional IRT model for a language placement test. Firstly, most placement tests have to measure language proficiency on a wide range of ability, from "absolute beginner" to "almost native". This requirement favours the development of testing strategies that minimize, during the administration, the number of items that are not relevant, either because they are too difficult or too easy. Computer adaptive testing, one of the most fruitful application of IRT is considered as a promising solution for placement testing (Larson and Madsen, 1985). Secondly, since very few language programs can afford to offer individualized instruction that would benefit from a diagnostic assessment based on various skills, the placement decision is usually unidimensional: the student is assigned a course which represents one step in a multi-level sequence. In many cases, the common axe for the placement decision is the general proficiency in the language.

Theoretical background and rationale

It seems that the debate raised by Oller (1979) on the structure of the language competence is still vigorous. Most theoretical models of competence assume that the communicative competence consists of various components (Canale and Swain, 1980; Bachman, 1990). For instance, Bachman claims that the communicative competence is a twofold construct: an organizational competence which can be further divided into a grammatical and a textual competence and a pragmatic competence which can be further divided into an illocutionary and a sociolinguistic competence. The development of these competencies rests on the development of various skills such as vocabulary, sensitivity to registers, cohesion marking etc. However, although many factors underlie different tests, many empirical studies have shown (Davidson, 1988; Henning and al.1985) that a first factor can explain an important proportion of variance. Whenever several factors are considered, the factor interpretation usually does not support consistently a given theoretical model (Vollmer 1983, Harley and al.,1987). According to Henning (1992), these results would suggest that there are two distinct states of dimensionality: psychological unidimensionality implies that test scores can be interpreted in regard of a known unitary construct whereas psychometric unidimensionality implies that the observed item variances are fairly homogeneous. Henning claims that psychometric unidimensionality is a sufficient condition to meet IRT requirements. Due to the way a second or foreign language is learnt one should expect underlying psychological dimensions of a test to be highly correlated (Carroll, 1987). This would mean that psychometric unidimensionality can be present on language tests even if they measure different skills.

However, instead of two different types of dimensionality, one could view the dimensionality of a test as a continuum. Because of the nature of human learning, all educational tests are likely, to some degree, to be multidimensional. There is some evidence that unidimensional IRT models such as the one used in this study are robust to the presence of a nondominant second dimension (Reckase 1979; Drasgow and Parsons, 1983; Doody-Bogan and Yen, 1983; Harrison, 1986; Blais, 1987). Consequently, meeting the requirements of unidimensional structure becomes a matter of showing that we are in fact dealing with a data

structure that is not too far from an appropriate unidimensional one and of highlighting important departures from unidimensionality.

Tests results could meet this basic requirement and still be so difficult to interpret that they could be totally meaningless. Therefore, a test that applies IRT procedure should be based on a sound theory and be used for decisions that are compatible with a unidimensional model. Any test that aims at measuring a general language proficiency trait should focus on basic skills that are correlated and that are activated in the current uses of the language. In addition, these instruments must be used for "unidimensional" decisions. Most placement tests, certification tests or selection tests belong to this category because they seek to locate a student on a single continuum, the general proficiency.

Assessing test dimensionality

Since the underlying "variables" of a test are not observable directly, unidimensionality must be demonstrated indirectly. The traditional psychometric approach to the assessment of dimensionality is through factor-analysis methods. In this approach, evidence of one dominant factor suggests that one single dimension adequately explains examinees' answers to all items. But in the past, the dimensionality of a test has also been assessed in a wide variety of ways. Hattie's (1985) list of various techniques for checking unidimensionality is impressive but the researcher found none of them fully satisfactory. These techniques may be grouped into four major categories:

- Indices based on response patterns: When ranked by their difficulty, items of a unidimensional test in relation with students' scores should provide a Guttman implicational scale. The degree to which the items fit the implicational scale can be estimated with various indices (Cliff, 1983). This type of evidence is also used by Henning (1992) to demonstrate language tests psychometric unidimensionality. In fact, these indices may be helpful to detect aberrant examinees or items patterns but cannot determine how much the test departs from unidimensionality and, if more than one, how many dimensions are present.
- Indices based on internal consistency: Classical item reliability coefficients (Cronbach alpha, KR-20, split-half index...) are usually low when a test measures different uncorrelated traits. At the item level, discrimination indices (biserial or point-biserial correlations and *a* parameters with 2-3 parameter IRT model) usually correspond to either poorly designed items or items which are not measuring the dominant trait.
- Indices based on principal component analysis or factor-analysis: Factor-analysis is usually made on the inter-item tetrachoric correlation matrix. It may be difficult to set a cut off proportion of the variance explained by the first factor under principal component analysis or to determine the number of factors that should be kept under factor-analysis. The eigenvalues may be plotted in order to determine if a dominant first factor is clearly present. Davidson (1988) used this technique to show that, on most common ESL standardized tests, a large proportion of item variance can be related to the first factor. Eigenvalues of the actual data may also be compared to eigenvalues obtained from an interitem correlation matrix of random data

(Drasgow and Lissak, 1983). Despite its mathematical complexity, fitting a nonlinear one-factor model as proposed by McDonald (1981) seems to be a very effective approach.

- Indices based on IRT calibration: Bejar (1980) proposed a procedure for investigating the unidimensionality of test by comparing item parameter estimates obtained on the full test and estimates obtained by dividing the test in subgroups, according to content. Hambleton and Rovinelli (1986) found that this method failed to recognize different dimensions on simulated data and suggested rather to use model fit index and compare expected values versus actual values of the residuals. Bejar recognizes the limitations of his method particularly when there is the same number of items in the subgroups but he argues that the method remains valuable as a descriptive tool and mentions that it should be used in conjunction with other procedures. As far as language tests are concerned, the Bejar procedure has been used to support language tests unidimensionality (Henning and al.,1985) but the method has also been questioned (Spurling 1987a, Henning 1987, Spurling 1987b).

More recently, to reflect the robustness of IRT estimations, the various degrees of dimensionality and the complexity of educational outcomes, Stout (1987) has proposed a non - parametric approach based on the concept of "essential unidimensionality". According to Stout, local independence of items should hold in subgroups of examinees of equal ability. The procedure he developed creates a set of assessment items that may load heavily on a second factor. The scores on the remaining items are used to partition the examinees into subgroups. A second set of assessment items is then created matching the difficulty with each of the items of the first assessment set. The variance estimates for each subgroup on the two assessment sets are then compared. A normalized t index is used to assess departure from unidimensionality. This method has been proven successful on rejecting the hypothesis of essential unidimensionality whenever the effect of the secondary dimension increases (Nandakumar, 1991). In language testing, Choi (1992) found this procedure more effective than Bejar's method of assessing the dimensionality of two widely used ESL tests, the *Test of English as a Foreign Language* (TOEFL) and the *University of Cambridge First Certificate of English* (FCE).

Method

The experimental (paper-and-pencil) version of the placement test consisted of 150 items (50 in each subtest). All the items were multiple-choice with four options. The test was administered to English-speaking Canadians enrolling in French summer classes in different colleges or universities. Most of these students were participating in a bursary programme for the learning of the second language in Canada. Due to the requirements of this programme, the population was fairly homogenous in terms of linguistic and cultural background and in terms of age, education and socio-economic status. However, since this study uses real data, some examinees who presented aberrant answer patterns were discarded to ensure the homogeneity of the sample. These examinees were detected using a reproducibility index obtained by creating a Guttman implicational scale of items and subjects.

On the first subtest (Paragraph), the student reads a paragraph of approximately thirty

words (for example, a short notice) and is asked a multiple-choice question. We suspect that this test basically rests on a single trait that could be labelled "reading". On the second subtest (Situation), the student reads the description, in English, of a current situation (for example, congratulating a friend) and the student must select the most appropriate statement among four French grammatically correct statements. The multiple aspects of appropriateness judgments suggest that the unidimensionality assumption may be violated in this subtest. The third subtest (Fill-the-gap) is a conventional "fill-the-gap" exercise which focuses on grammar use and vocabulary. These two components may represent two dominant factors.

The first data set consisted of the answers of 348 students who had completed the whole test. The data was then augmented with the answers of students who had written only one or two subtests. The scores on each subtest were regressed to make sure that this larger data set had the same distribution as the original one. After deleting some examinees we could create a sample of 694 examinees for the first subtest (Paragraph), 681 for the second subtest (Situation) and 661 for the third subtest (Fill-the-gap).

Four different approaches were used to determine how the items depart from the unidimensionality assumption:

- A structural equation approach: We tried to fit a LISREL (Jöreskog and Sörbom, 1983) model that would confirm the relevance of the three-subtest division. Each subtest was divided in two equal parts: items with even numbers and items with odd numbers. A single factor solution and a three factor one (for each subtest) were tested.
- A factor-analysis approach: Using TESTFACT (Wilson and al.,1991), we examined the factor structure of each subtest in order to determine the role of the first factor and to identify major clusters of items.
- A non-parametric approach: We ran the DIMTEST (Stout and al., 1991) program to verify the essential unidimensionality of each subtest. The grouping of the items was first based on a factor-analysis item grouping and then on an expert content grouping.
- An IRT based approach: According to a content analysis, the items in each subtest were grouped in two sets. Using Bejar's method, we compared the estimation of the IRT parameters obtained with BILOG (Bock and Aitken, 1981; Mislevy and Bock, 1986). The difficulty parameters were then plotted and the regression slope was used as a statistic.

Results

1) The structural equation approach

A LISREL model assumes that there is a causal structure among a set of latent and observed variables. The program can be useful to test various hypothesis regarding the language competence underlying students answers (Sang and al.,1986). We intended to check if a simple model could be fitted with the answers from the student who had completed the whole test (n=348). We divided each subtest in two equal parts of 25 items: these parts were created randomly, keeping the items with an even number in one part and those with an odd number in

the other part. The correlation between the two parts of a subtest can be interpreted as a reliability index (split-half index). As shown in table 1, these correlations are fairly high. In addition, the correlations involving different subtests were also so high that we could believe a single factor model could be acceptable.

	#1: Paragraph		#2: Situation		#3: Fill-the-gap	
	ODD	EVEN	ODD	EVEN	ODD	EVEN
#1 ODD	1.000	0.996	0.857	0.857	0.873	0.873
#1 EVEN	0.996	1.000	0.853	0.853	0.873	0.873
#2 ODD	0.857	0.853	1.000	0.992	0.809	0.811
#2 EVEN	0.857	0.853	0.992	1.000	0.805	0.808
#3 ODD	0.873	0.873	0.809	0.805	1.000	0.995
#3 EVEN	0.873	0.873	0.811	0.808	0.995	1.000

Table 1: *Correlation between subtests and parts*

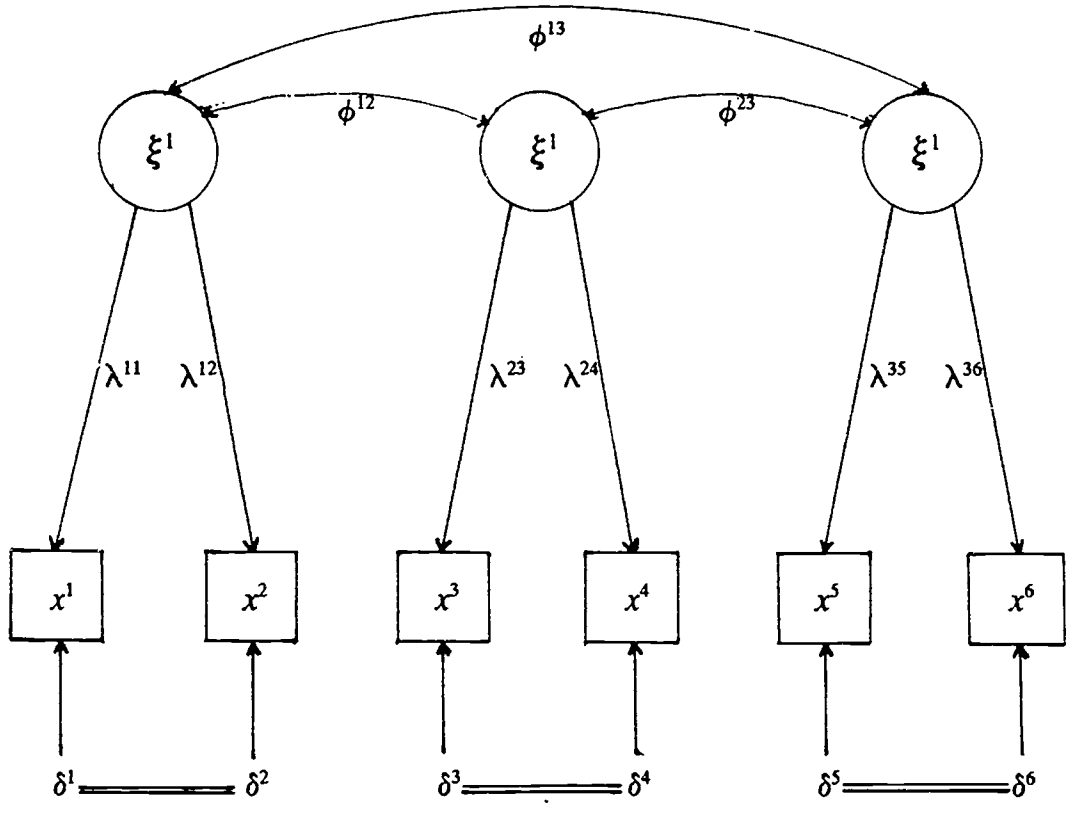


Figure 1 : *3-factor LISREL model*

However, since there was a significant amount of residuals this single factor model provided a poor fit. The best fit was obtained with a three correlated-factor solution. This solution is illustrated in figure 1. The squares corresponds to the observed variables (x), the examinees' scores on each part. These scores consist of a measurement error (δ) and the effect of one of the latent variables. The circles (ξ) corresponds to the latent variables underlying the observed variables. The relation between the observed and latent variables is identified as λ. Thus, each factor is related to a single subtest. This support the decision of separating the items in three different banks. The symbol ϕ represents the correlation between the factors. These correlations are fairly high as they range from .81 to 87. This oblique solution is certainly not the most attractive and the most explanatory but is theoreticaly tenable for a general proficiency language test. These results alone do not provide sufficient evidence regarding the unidimensionality of each subtest. However, if any of the three subtests measures more than one trait, the LISREL analysis shows that the internal consistency is clearly not affected.

2) The factor-analysis approach

There are some problems related to the use of traditional factor-analysis methods with dichotomous variables (i.e., items scored right or wrong). The classical factor-analysis model does not fit the dichotomously scored test items. This difficulty arises from the fact that the common factor variables, and also the unique factor variables, are usually conceived as taking continuous values. In general, of course, it is a contradiction to consider a discrete-valued random variable to be a linear combination of a set of continuous variables (Lord and Novick, 1968:357). Some problems also arise when using tetrachoric correlations in item factor-analysis (Green and al.,1984). Fortunately, item factor-analysis strategies have been developed to deal with these difficulties (Mislevy, 1986). One such method, "full information factor-analysis" developed by Bock and al. (1985) was used in this study to explore the dimensionality of examinees' responses.

Full information factor-analysis is a marginal maximum likelihood procedure that provides a test of the dimensionality of the underlying ability continuum, and, when multidimensionality is found, it may indicate how the items may be partitioned into unidimensional sets (Bock, 1984). Because the interitem correlations are not used in the computations, the full information approach avoids problems associated with interitem correlations leading to the emergence of "difficulty" factors. The methods proceed similarly as structural equation modelling; a comparison is made of the likelihood ratio goodness-of-fit G^2 tests obtained for models including a varying number of factors. Technical details of the full information factor-analysis method are presented in Annex A.

As it is mentioned by Steinberg and al.(1990:214), although this method provides a statistical determination of the dimensionality of the items in a test, discerning the content of each factor is done by inspection of the factor loadings. When the pattern of factor loadings provides a meaningful division of items, the item pool may be partitioned into unidimensional subtests. Under some circumstances, there may be evidence of multidimensionality that is not accompanied by any clear indication in the pattern of factor loadings, of how the items may be

partitioned to achieve unidimensionality. Zimowski and Bock (1987) suggest that partitioning the test into unidimensional subtests should be done only when there is clear distinction between the cognitive skills involved and when these skills correspond to distinctions between the areas in which the test is to predict performance.

The full information technique was applied to the total test and all three subtests. The results presented in Annex B suggest to us the following interpretation. When applied to the whole test (150 items, n=348), the TESTFACT 3-factor analysis incicates that the first factor explains 25.4% of the variance. Adding a second and a third factor allows to explain an additional 2.4% and 1.4%. As expected, the correlation between the factors are fairly high: .6 between the first and the second factor, .68 between the first and the third one, and .49 between the second and the third one. This indicates that the oblique (PROMAX) analysis is preferable. The inspection of loadings on each factor shows that the first 50 items (Subtest #1, Paragraph) are mostly related to the first factor that could be labeled as a reading factor. Reading would then be a important facet of the general proficiency. This factor is also important with the last 50 items (Subtest #3, Fill-the-gap) although most of these items are also clearly related to the third factor. Therefore, we can conclude that the last subtest involves a reading ability in the second language and a specific ability which is associated with grammar, vocabulary and a particular "Fill-the-gap" skill. The second factor is more difficult to interpret. The few items that load heavily on this factor are generally items that involve some logical reasoning which may not be related to language abilities. As far as Subtest #2 (Situation) is concerned, there is not a clear pattern of factor loading. However, the loadings are usually lower which could suggest that this subtest is less reliable and/or measures numerous abilities. This is not surprising since the jugdments on the appropriateness of a statement in a given situation may refer to a different knowledge of the world, a different cultural background, a different perception of the described situation etc. At any rate, the distribution of loadings suggests that the three suotests are measuring different constructs and should be separated.

Running TESTFACT on each subtest gives a different perspective. With 33% of the variance explained with the first factor, on the 3-factor-analysis, Subtest #1 (Paragraph) is the most strongly related to a dominant factor. The proportion is even higher with the 2-factor solution as the proportion explained by the first factor increases to 34% of the variance whereas the proportion on the second one drops to 2.3%. This suggests this subtest is fairly unidimensional. The oblique analysis gives a correlation between these two factors of .7. Looking at the loadings on these two factors, one may conclude that they tap two different although correlated cognitive abilities: the first factor involves the ability to reformulate an information in the second language and the second factor seems to involve the ability to draw inferences from an information.

The proportion of explained variance is much lower for the second test (Situation). Even with three factors, the sum of explined variance is only 23.8%. Only six items show a loading on the first factor higher than .5. These six items are related to the ability to distinguish different registers (formal vs. familiar) although other items where this ability should be present do not load as heavily as on the first factor. Noteworthy is the fact that 11 items show a loading

smaller than .3 on all three factors. One can conclude it is very difficult to devise a unidimensional test that consists of items that measure the abilty to select the most appropriate statement in a given communicative situation.

As far as Subtest #3 (Fill-the-gap) is concerned, since the task is not as integrative as on the first two subtests, we had good reasons to believe that there would be a division between items focusing on grammatical discrete points and those focusing on vocabulary discrete points. Inspection of the loading on the 2-factor solution does not support such a distinction although right answers on the items strongly associated with the first factor assume a correct semantic interpretation of the sentence. On the other hand, the items related to the second factor focus on the application of formal rules (for example, the pronominalization transformation rule). The TESTFACT results show that the 3-factor solution explains more variance than the 2-factor solution by inflating the first factor role (from 23.3% to 28.3%).

3) The non-parametric approach

Stout (1987) has developed a non-parametric statistical procedure (DIMTEST) based on the large sample distribution theory for assessing latent trait dimensionality and has argued the validity of this procedure based on simulation studies involving a wide variety of achievement tests. Stout provides a definition of the number of dominant dimensions known as essential dimensionality. Briefly we can say that the essential dimensionality (d_E) of an item pool U is the minimal dimensionality (number of elements in the ability vector θ) necessary to satisfy the assumption of essential independence. Essential independence implies that the average value of the item's covariance given the vector of ability θ over all item pairs is small in magnitude for all θ as the test length increases. When $d_E = 1$, essential unidimensionality is said to hold.

The statistical procedure for testing the null hypothesis of essential dimensionality is described in detail in Annex A. The idea is to split the original set of items into partitioning subtests on which comparative computing will be made. The N test items are split into two assessment subtests of length M each, called the Assessment 1 subtest (AT1) and the Assessment 2 subtest (AT2), and a longer subtest, called the partitioning subtest (PT) of length n = N - 2M. The M items for subtest AT1 are selected to have the same dominant trait. This splitting can be done using either expert opinion or exploratory factor-analysis. Once items for AT1 are selected, a second set of M items for AT2 is selected from the remaining items so that AT2 items have a difficulty distribution similar to AT1 items. The remaining n items then become the partitioning subtest PT.

Each examinee is then assigned to one of K subgroups according to his score on PT. After eliminating subgroups with too few examinees, within each subgroup k, two variances estimates, the usual variance estimate, and the unidimensional variance estimate are computed using items of AT1. The difference in these variance estimates is then normalized by an appropriate normalizing constant and summed over subgroup to arrive at the statistic T_L. Similarly, using items of AT2, the two variance estimates and the standard error of estimate are computed and normalized within each group to arrive at the statistic T_B. The statistic T, to assess

departure from essential unidimensional is then calculated using T_L and T_B. The null hypothesis of d_E is rejected if $T \geq Z_\alpha$, where Z_α is the upper $100(1-\alpha)$ percentile of the standard distribution and α is the desired level of significance.

All three subtests were submitted to an essential unidimensionality study. The main results are presented in Annex B. We report hereafter the conclusions of the three analyses we ran with DIMTEST on each subtest[2]. One strategy in essential dimensionality assessment consists in asking the program to select the items in the assessment set (AT1) according to the results of the factor analysis. Under this strategy, the program will select items loading on the second factor. As shown in Table 2, it seems that this strategy is less sensitive to departure in unidimensionality. That is to say that the program failed to recognize more than one dimension in any of the three subtests.

	Subtest #1 Paragraph	Subtest #2 Situation	Subtest # 3 Fill-the-gap
2nd factor	-1.497	.279	-.789
Domain A	.78	1.872*	1.707*
Domain B	.589	1.641*	2.76**

* Significant at $\alpha=.05$ ** Significant at $\alpha=.005$

Table 2: *Values of T obtained from DIMTEST*

The other strategy that is available consists in selecting the items according to the judgment of an expert upon the content of the items. The expert, a graduate student in second language acquisition[3], was asked to classify the 50 items of each subtest in two or three major domains. Most of the items fall in one of two categories:

- Subtest #1, Paragraph — Domain A: *Reformulating ability* (18 items); Domain B: *Inference ability* (28 items)
- Subtest #2, Situation: — Domain A: *Lexical competence* (29 items); Domain B: *Sociolinguistic competence* (21 items)
- Subtest #3, Fill-the-gap: — Domain A: *Vocabulary knowledge* (22 items); Domain B: *Grammar rules* (28 items)

We selected 12 items in each domain to create two different assessment sets (AT1). The number of items is within the limits recommanded by Nandakumar and Stout (1993). It also allows to cover the full ablility range. The results suggest that this expert strategy is more appropriate in order to detect departure from unidimensionality. Under this strategy the Subtest #1 (Paragraph) was found unidimensional. The unidimensionality hypothesis was rejected on the other two subtests at a $\alpha=.05$ significance level. However only the grammar component on Subtest #3 (Fill-the-gap) was clearly a distinct dimension as the null hypothesis was rejected at a $\alpha=.005$ level. This approach leads us to consider the two cognitive tasks of Subtest #1

[2] We are grateful to Dr. William Stout who generously provided the DIMTEST program.

[3] We would like to thank Mee Liam Chung-How for her important contribution in this research.

(Reformulating and Inference) as a single trait but grammar and vocabulary as two distinct traits measured by Subtest #3. The portrait is not so clear as far as Subtest #2 (Situation) is concerned. This may be due to the fact that the two competencies (lexical and sociolinguistic) are not unitary constructs.

4) The IRT-Based approach

The item banks that are being scrutinized are used in conjunction with an adaptive testing strategy based on IRT principles. During an adaptive test, the selection of the items depends on the current estimation of the student's ability so that final placement will be based on different sets of items. This is possible because of the IRT invariance properties. It is assumed that the ability estimates should not differ significantly whenever answers to different items are used. It is also assumed that the item parameters are constant, when different combinations of items are used for the calibration.

The procedure proposed by Bejar (1980) calls for obtaining two sets of item parameter estimates. In our particular case one set is obtained from the initial calibration including the 50 items of each subtest; the other set is obtained by doing a calibration using only the items within a content area. The expert division used for the essential dimensionality approach was kept for the second calibration so that each subtest leads to two comparisons with the initial parameter estimates: Domain A vs. total subtest and Domain B vs. total subtest. Bejar's rationale is that both sets of parameters should be linearly related (within estimation errors) unless one or more content area is tapping a dimension which is unique.

If unidimensionality holds, the plot of the content-area parameters and the whole subtest parameters should illustrate a perfect correlation, dots being along a theoretical axis with a slope of 1.00 an intercept of 0.0. In fact, even though a 3-parameter model was used, only the difficulty estimates (parameter *b*) were plotted. These values are more informative because discrimination (parameter *a*) and pseudo-guessing (parameter *c*) are usually not estimated as accurately as difficulty parameters.

Bejar proposes two criteria for judging unidimensionality. One is the comparison of the actual axe with the theoretical one. Classical linear regression procedures can be applied to calculate the slope and the intercept. According to Bejar, differences in intercepts would mean that the two sets differ only by a constant which is an artifact of the estimation program. Because of the indeterminacy of the origin of the difficulty and ability scale under IRT, the 0 point is usually set to the mean difficulty of the items. However it should be kept in mind that 3-parameter models place difficulty on a logistic scale; this means that, unless the origin is common, variations at the extreme end of the scale may affect the slope. Differences in slope are more serious because they can be interpreted as differences due to content effect at different levels of ability. The second criterion examines the mean distance of items within the different content-area. In other words this criterion refers to the residuals of the regression analysis. Items whose difficulty values strongly diverge, should be considered as source of departure from unidimensionality.

Discussion

When we began this research, we believed that the relevance of the threefold division would be confirmed. The structural equation approach (LISREL) results and factor-analysis approach indicated that the division should be maintained. TESTFACT indicated that each subtest included more than one dimension although it has been difficult to relate the factors to traits that would be components of a model of communicative competence. This was particularly difficult for the Subtest #2 (Situation), which construct seems to be extremely complex. This kind of content based on appropriateness judgments certainly belongs to communicative competence but may require a bias study (like a DIF study) to complement the present study. The results for Subtest #3 (Fill-the-gap) did not fully support the distinction between grammar and vocabulary but revealed a distinction between items involving semantic interpretation and items requiring the application of formal rules. Subtest #1 (Paragraph) is the most unidimensional part with a dominant factor that could be labelled "Reading". The essential dimensionality approach, based on expert judgment, rejected the multidimensionality hypothesis for this subtest whereas results were on the borderline as far as the other two subtests are concerned. The Bejar's approach overlooked these problems with Subtests #2 and #3 but located some suspicious items in Subtest #1.

We also found that within a given approach, different analyses could lead to diverging results. For example, DIMTEST produces different results when expert judgment is used as an input instead of exploratory factor analysis. Different results are obtained with TESTFACT if a 2-factor rather than 3-factor analysis is called. LISREL analysis would certainly have fit a different model if the division between items have been based on content rather than on an arbitrary odd/even distinction. Bejar's method could eventually yield an almost perfect correlation between content-based and total subtest if outlying items were discarded.

Careful attention should be paid to the robustness of these different approaches. As far as factor-analysis approaches are concerned, there is a consensus regarding the role of the first factor as an indication of unidimensionality. Factor loadings may be inspected in order to detect additional traits that could be measured or to delete items that are not measuring the dominant trait. Regarding the non-parametric approach, sensitivity to departure from unidimensionality has to be more documented. The default factor analysis seems to select items that are related to a difficulty factor rather than an interpretable trait. There are also problems regarding the numbers of items to be included in the assessment set. If the number of items related to a content exceeds N/4, how should these additional items be considered? Bejar's method has been criticized because it fails to recognize various dimensions in a test. Plotting the ability estimates rather than the difficulty estimates as suggested by Hambleton and Rovinelli (1986) may be a valuable method if an application calls for invariant estimates of ability as in adaptive testing.

On the basis of the four analyses that have been conducted, one may question the concept of unidimensionality. Various approaches seem to lead to different and sometimes conflicting results. These findings suggest that unidimensionality is not itself a unidimensional concept at all! The concept is defined through the approach. The issue that is now raised is not simply

The parameters were estimated by BILOG program. This program is particularly suited for this kind of analysis because it provides a standard error of estimate for each parameter. The program was ran once on the 50 items of each subtest and then twice keeping only the items that were assigned to Domain A or B. Scattergrams are given in Annex B.

	Correlation r	Slope β	Intercept α	Mean square
Subtest #1 (Paragraph)				
A: *Reformulating ability*	.998	.972	.051	12.62
B: *Inference ability*	.657	.641	-.0933	14.316
Subtest #2 (Situation)				
A: *Lexical competence*	.999	.988	0	41.647
B: *Sociolinguistic competence*	.896	1.017	.204	45.086
Subtest #3 (Fill-the-gap)				
A: *Vocabulary knowledge*	.97	1.031	.130	41.491
B: *Grammar rules*	.984	1.039	-.057	28.157

Table 3: *Total subtest vs. Content-based parameter estimates*

The intercept and slope estimates were computed by SPSS-PC (REGRESSION) program. Except for Subtest #1 (Paragraph), the slope are close to 1. In this regard, the procedure did not detect any departure from unidimensionality for Subtest #2 (Situation) and Subtest #3 (Fill-the-gap). The slope for Domain A of Subtest #1 (inference ability) is only .641. From this point of view, the IRT-based approach detected some problems with Subtest #1 (Paragraph). These problems were ignored by TESTFACT and DIMTEST. In addition, BILOG goodness-of-fit statistics did not show any sign of misfit for these items.

Since difficulty parameters are placed on a logistic scale, most important discrepancies are likely to occur at both ends of the difficulty range. That is the reason why most of the outlying items located by SPSS-PC were easy or very difficult items. The inspection of the scattergrams created under the IRT-based approach clearly shows that four items (all related to the inference ability) in Subtest #1 were given very different difficulty values. Whether or not, these items measure a specific trait may be a controversial issue. However, since the parameters are obviously not invariant, these items should be excluded from the bank. From this point of view, the procedure can be a valuable complement to the item analysis.

which approach is the best one but what kind of unidimensionality we look for. Item analysis based on the discrimination indices may be sufficient in small-scale testing provided that the scope of the test is clearly defined. Estimations of IRT item parameters call for analyses that aim at determining if a general (albeit conceptually ambiguous) factor emerges or if the factors are strongly correlated (Harrison, 1986). From this point of view the structural-equation approach combined with the IRT-based approach could probably ensure an accurate estimation of the parameters. Nonetheless, the question of using these parameters to design a placement test must be addressed with a factor analysis or an essential dimensionality study in order to assess the capability of this instrument to measure a general proficiency trait. A conjunction of various approaches may even help to validate language tests with respect of theoretical models.

Unidimensionality is not a yes/no issue; it is rather a matter of degree considering the purpose of the test. To what extent a departure from unidimensionality rules out the use of a test in a given situation? In fact we know that a language test is never fully unidimensional. Moreover, unidimensionality concerns should not force test developers to restrict the nature and the range of tasks whenever validity would entail diversity and complexity. Dimensionality analysis as part of the construct validity study of a test is a long term process, in search of an adequate but perfectible construct.

Finally this study shows how important is the expert judgment in assessing unidimensionality. For example, DIMTEST takes advantage of an expert-based item grouping. TESTFACT results, as any other factor analysis results are meaningless if they are not interpreted in the light of a sound theory though the expert's eyes. Structural equation modelling should be done cautiously since it may lead to peculiar results if it is not based on a sound theory of communicative competence. However the importance of expert judgment should not be a reason for rejecting more empirical approaches. What the expert thinks must be supported with some evidence. On the other hand, any empirical analysis must be driven, interpreted and confirmed by an informed judgment.

References

Bachman, L.F. 1990: *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bejar, I.I. 1980: A procedure for investigating the unidimensionality of achievement tests based on item parameters estimates.*Journal of Educational Measurement* 17, 283-96.

_____1988: An approach to assessing unidimensionality revisited. *Applied Psychological Measurement* 12, 377-79.

Blais, J.-G. 1987: Effets de la violation du postulat d'unidimensionalité dans la théorie des réponses aux items. Unpublished doctoral thesis. Faculté des sciences de l'éducation. Université de Montréal. 181 pages.

Bock, R.D. 1984: Full information factor analysis. Paper presented at the annual meeting of the psychometric society, Santa Barbara, CA.

Bock, R.D. and Aitkin, M. 1981: Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* 46, 443-459.

Bock, R.D., Gibbons, R.D., and Muraki, E. 1985: *Full information factor analysis*. (MRC Report No 85-1). Chicago: National Opinion Research Center.

Bock, R.D. and Lieberman, M. 1970: Fitting a response model for n dichotomously scored items. *Psychometrika*, 179-197.

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

Carroll, J.B. 1983: Psychometric theory and language testing. In Crotjahn R., Klein-Braley C. and Stevenson D.K., *Taking Their Measure: The Validity and Validation of Language Tests*. Bochum: Studienverlag Dr. N, Brocknegen, 1-40.

Choi, I.C. and Bachman, L.F. 1992: An investigation into the adequacy of three IRT models for data from two EFL reading tests. *Language Testing* 9, 51-78.

Cliff, N. 1983: Evaluating Guttman Scales: some old and new thoughts. In Howard Wainer and Samuel Messick (Ed.), *Principals of Modern Psychological Measurement*. Hillsdale N.T., Laurence-Erlbaum, 283-301

Davidson, F. 1988: An explanatory modeling survey of the trait structures of some existing language test data sets. Unpublished Ph.D dissertation. Department of Applied Linguistics, University of California, Los Angeles.

Doody-Bogan, E. and Yen, W.M. 1983: Detecting muitidimensionality and examinig its effect on vertical equating with the three paramater logistic model. Paper presented at the annual meeting of the American Educational Research Association, Montreal, April 1983.

Drasgow, F. and Lissak, R.I. 1983: Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology* 68, 363-73.

Drasgow, F. and Parsons, C.K. 1983: Application of unidimensional psychological item response theory models to muitidimensional data. *Applied Psychological Measurement* 7, 189-99.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L. and Reckase, M.D. 1984: Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21, 347-360.

Haberman, J.S. 1977: Log linear models and frequency tables with small expected all counts. *Annals of Statistics*, 5, 1141-1169.

Hambleton, R.K. and Rovinelli, R.J. 1986: Assessing the dimensionality of a set of test items. *Applied Psychological Measurement* 10, 287-302.

Harley, B., Allen, J.B., Cummins, J. and Swain, M. 1987: *The development of bilingual proficiency : final report*. Toronto: Modern Language Centre, Ontario Institute for Studies in Education.

Harrison, D.A. 1986: Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, Vol.11, no.2, 91-115.

Hattie, J.A. 1984: An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research* 19, 49-78

_____1985: Methodology review: Assessing unidimensionality ot test and items. *Applied Psychological Measurement* 9, 139-64

Henning, G. 1987b: Is the Bejar test of unidimensionality appropriate? A response to Spurling. *Language Testing* 4, 96-98.

_____1992a: Dimensionality and construct validity of language tests. *Language Testing* 9, 1-11.

_____1992b: Dimensionality and construct validity of language tests. *Paper presented at the 14th Language Testing Research Colloquium*, Vancouver 1992.

Henning, G.T., Hudson, T. and Turner, J. 1985: Item response theory and the assumption of unidimensionality. *Language Testing* 2, 141-54.

Jöreskog, K.W. and Sörbom, D. 1983: *LISREL User's guide.* Department of Statistics, University of Uppsala.

Larson, J.W. and Madsen, M.S. 1985: Computerized adaptive language testing. Moving beyond computer-assisted testing. *CALICO Journal*, 32-43.

Laurier, Michel 1991: What we can do with computerized adaptive testing... And what we cannot do! *Currents Developments in Language Testing*, Anthology Series 25, 244-255. SEAMEO Regional Language Centre.

Lord, F.M. and Novick, M.R. 1968: *Statistical theories of mental test scores.* Reading Mass: Addison-Wesley.

McDonald, R.P. 1981: The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology* 27, 82-89.

Mislevy, R.J. 1986: Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics* 11, 3-31.

Mislevy, R.J. and Bock, R.D. 1986: *PC-BILOG.* Chicago, IL: Scientific Software Inc.

Nandakumar, R. 1991: Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement* 28, No.2, 99-117.

Nandakumar, R. and Stout, W. 1993: Refinement of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics* 18, 41-68.

Oller, J.W.Jr. 1979: *Language Tests at School.* London: Longman.

Reckase, M.D. 1979: Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics* 4, 207-230.

Reckase, M.D., Ackerman, T.A. and Carlson, J.E. 1988: Building unidimensional tests using multidimensional items. *Journal of Educational Measurement* 25, 193-203.

Sang, F., Schmitz,H.J., Vollmer, H.J., Baumert, J. and Roeder, P.M. 1986: Models of second language competence: a structural equation approach. *Language Testing* 3. 54-79.

Spurling, S. 1987a: Questioning the use of the Bejar method to determine unidimensionality. *Language Testing* 4, 93-95.

_____1987b: The Bejar method with an example: a comment on Henning's "Response to Spurling". *Language Testing* 4, 221-24.

Steinberg, L., Thissen, D., and Wainer, H. 1990: Validity. In H. Wainer (ed), *Computerized Adaptive Testing: A primer.* Hillsdale, N.J.: Erlbaum.

Stout, W. 1987: A non-parametric approach for assessing latent trait unidimensionality. *Psychometrika* 52, 589-617.

_____1991: *DIMTEST and TESTSIM.* University of Illinois, Urbana-Champain.IL.

Traub, R.E. 1983: A priori considerations in choosing an item response model. In R.K. Hambleton (Ed.), *Applications of item response theory* , 57-70. British Columbia, Canada: Educational Research Institute of British Columbia.

Vollmer, H. 1983: The structure of foreign language competence. In Arthur Hayles and Don Porter: *Current Development in Language Testing.* London: Academic Press, 3-29.

Wilson, D., Wood, R.L. and Gibbons, R.D. 1991: *Testfact: Testscoring and item factor analysis* (Computer program). Chicago: Scientific Software.

Zimowski, M.F. and Bock, R.D. 1987: *Full information item factor analysis of test forms from the ASVAB CAT pool.*(MRC Report No 87-1), Chicago, IL: Methodology Research Center, National Opinion Research Center.

Annex A

Full information factor-analysis approach

The full information factor analysis was implemented by using the TESTFACT program (Wilson, Wood, Gibbons, 1983). This approach uses examinees' item response vectors instead of correlations between items.

In applying this method, a specific model for the item responses must be assumed. The selected model was a multivariate generalization of the three-parameter normal ogive in which each item is allowed to load on multiple factors. The model can be developed by first assuming that underlying the response of person i on item j is a response variable defined as:

$$y_{ij} = \sum_{k=1}^{K} \lambda_{jk}\theta_{ki} + \nu_j$$

where θ_{ki} represents the value of the k^{th} latent variable (factor), k= 1, 2,...K for the i^{th} individual, i=1,2,...,N; λ_{jk} is the loading of the j^{th} item, j=1,2,...,n, on the k^{th} latent variable, and υ_j is a residual term associated with item j.

The response variables are assumed to have mean zero and variance one. The observed score of the i^{th} examinee on the j^{th} item, x_{ij}, takes on a value of one, indicating a correct score, if y_{ij} exceeds g_j, the threshold for the j^{th} item. Otherwise, $x_{ij} = 0$. If it is assumed that the residuals n_j are independently distributed as N(0, s_j) , the conditional probability that the i^{th} examinee gets the j^{th} item correct, given that examinee's values on the latent variables are equal to the vector $\theta_i = \{\theta_{1i}, \theta_{2i}, \ldots, \theta_{Ki}\}$ can be expressed as:

$$P(x_{ij} = 1 \mid \theta_i) = \frac{1}{\sqrt{2\Pi}\sigma_j} \int_{\gamma_j}^{\infty} \exp\left[-1/2 \left(\frac{y_{ij} - \sum_{k=1}^{K} \lambda_{jk}\theta_{ki}}{\sigma_j} \right)^2 \right] dy$$

$$= F_j(\theta)$$

This is a multivariate generalization of the two-parameter normal ogive model (see Lord & Novick, 1968, Chapter 15). This model can be modified to allow for the possibility of guessing by substituting:

$$F_j^*(\theta_i) = c_j + (1 - c_j)F_j(\theta_i)$$

for $F_j(\theta_i)$, where c_j represents the probability that an individual with very low ability gets the item correct. The c_j values are treated as fixed constants in the full information factor-analysis. In this study, the c_j were fixed to the inverse of the number of choices for multiple-choice items. These fixed values served as input values to the TESTFACT program. Omitted items were scored as wrong answers.

Incorporating the item response function, $F^*_j(\theta_i)$, defined in a previous equation, the marginal probability of the s^{th} response pattern can be expressed as:

$$P_s = P(\mathbf{x}=\mathbf{x}_s) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j=1}^{n} F^*_j(\theta_j)^{x_{sj}} [1 - F^*_j(\theta_j)]^{1-x_{sj}} f(\theta)\, d(\theta)$$

where x_{sj} is the response to the j^{th} item in the s^{th} response pattern, $s = 1,2,\ldots,S$, and $S \leq \min(2^n, N)$ is the number of response patterns. It is further assumed in this application that $f(\theta)$ is the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_x$. If it is assumed that the counts of the distinct response patterns follow a multinominal distribution, the likelihood of the matrix $\mathbf{X}$ of observed counts r_s of distinct response patterns can be expressed as:

$$P(\mathbf{X}) = \frac{N!}{r_1!r_2!\cdots r_s!} P_1^{r_1} P_2^{r_2} \cdots P_s^{r_s}$$

where P_s is given by the previous equation.

The quantities P_s are approximated using numerical integration. The marginal maximum likelihood method of Bock and Aitkin (1981), which is based on earlier work by Bock and Lieberman (1970) is then applied to the previous equation to obtain estimates of the factor loadings and thresholds for each item (see Bock and al., 1985; Mislevy, 1986).

If sample size is sufficiently large, a test of the fit of the K-factor model relative to a general multinominal alternative can be obtained using a chi-square approximation to the likelihood ratio test. The model can be re-estimated and the test repeated for successive values of K. The difference between these chi-square statistics is also distributed as chi-square and can be used to test the improvement in model fit that is achieved by allowing the number of latent variables to increase. The test of change in model fit has been shown to perform well even when the frequency table is sparse (Haberman, 1977).

Essential dimensionality approach

The procedure described by Stout (1987) and ameliorated by Nandakumar and Stout (1993) uses a statistic T to test the hypothesis that there is only one dimension, $H_0 : d=1$, versus the alternative that there is more than one dimension $H_1 : d>1$.

Observations (raw data) can be represented by $\{U_{ij}\}$ where i, $1 \leq i \leq n+M$, indexes items and j, $1 \leq j \leq J$, indexes examinees. In the present case, observations are response vectors of 0's and 1's with 1 denoting a correct response to an item and 0 denoting an incorrect response. The different steps to calculate T when the number of examinees is "small" (under 2000) are the following:

Step 1: Split test into partitioning and assessment subtests.

The N test items are split into a short assessment subtest of length M and a long partitioning subtest of length n. For some robustness considerations it is preferable that $4 \leq M \leq N/4$ (Nandakumar and Stout, 1991).

The M items can be chosen along two strategies. They represent a homogeneous set of items in the opinion of an expert or they load most heavily positively or negatively on the second extracted factor of a principal axis factor analysis (with no rotation) of the tetrachoric correlation coefficients with maximum observed correlations in each column used in place of communalities.

Step 2: Assign examinees to subgroups.

The examinees are assigned to different subgroups according to their differing partitioning subtest scores. It is required that each subgroup has a "large" number of examinees. All subgroups with less than J_{min} examinees are deleted and $J_{min} \geq 20$ is recommended to maintain close agreement with the asymptotic theory.

Step 3: Compute the "usual" variance estimate for the k-th subgroup.

Let U_{ijk} indicate the correctness of the response of the j-th examinee from subgroup k to the i-th assessment item.
Let $J_k \equiv J_k^{(n)}$ denote the number of examinees of subgroup k and $K \equiv K^{(n)} \leq n-1$ denote the number of subgroups.
Let

$$Y_j^{(k)} = \sum_{i=1}^{M} \frac{U_{ijk}}{M}$$

the assessment subtest score of the j-th examinee from subgroup k.

Let

$$\bar{Y}^{(k)} = \sum_{j=1}^{J_k} \frac{Y_j^{(k)}}{J_k}$$

the average examinee assessment subtest score for subgroup k.

Let

$$\hat{\sigma}_k^2 = \sum_{j=1}^{J_k} \frac{(Y_j^{(k)} - \bar{Y}^{(k)})^2}{J_k}$$

the usual variance estimate of examinee assessment subtest scores in subgroup k.

Step 4: Compute the unidimensional variance estimate for the k-th subgroup

Let

$$\hat{p}_i^{(k)} = \sum_{j=1}^{J_k} \frac{U_{ijk}}{J_k}$$

and

$$\hat{\sigma}_{U,k}^2 = \sum_{i=1}^{M} \frac{\hat{p}_i^{(k)}(1 - \hat{p}_i^{(k)})}{M^2}$$

the unidimensional variance estimate for subgroup k.

Step 5: Normalize and combine the different subgroup variance estimates to form the statistic

Let

$$\hat{\mu}_{4,k} = \sum_{j=1}^{J_k} \frac{(Y_j^{(k)} - \bar{Y}^{(k)})^4}{J_k}$$

$$\hat{\delta}_{4,k} = \sum_{i=1}^{M} \hat{p}_i^{(k)}(1 - \hat{p}_i^{(k)})(1 - 2\hat{p}_i^{(k)})^2$$

and

$$S_k'^2 = \left[\left(\hat{\mu}_{4,k} - \hat{\sigma}_k^4\right) + \hat{\delta}_{4,k}/M^4\right] / J_k$$

Now let the statistic T_L be:

$$T_L = \frac{1}{K^{1/2}} \left(\sum_{k=1}^{K} \frac{X_k}{S_k'} \right)$$

where:

$$X_k = \hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2$$

Each X_k measures non-unidimensionality in the sense that $X_k \equiv 0$ when $d_E = 1$ and $X_k > 0$ on average when $d_E > 1$.

Step 6: Correct for statistical bias

Select a set of M items from what would otherwise be the n partitioning subtest items such that the selected items have an item difficulty distribution as similar as possible to that of the assessment subtest. Compute T_B like T_L, but using the assessment subtest 2 items. the bias corrected statistic is defined by:

$$T = \frac{(T_L - T_B)}{2^{1/2}}$$

Since assessment subtest are selected to have a difficulty distribution similar to that of assessment subtest 1, this allows T_B to compensate for the influence of item difficulty bias on T_L. Also T_B simultaneously compensates for the influence of examinee variability bias on T_L.

Step 7: Perform the test for unidimensionality for J small; that is, $J \leq 2000$

Reject H_0: $d = 1$ if $T > Z_\alpha$, where Z_α is the upper $100(1-\alpha)$ percentile for a standard normal distribution, α being the desired level of significance.

Annex B

Results of full-information factor-analysis approach

Total test

Number of common factors = 2

Chi-square = 43362.90 DF = 136
Change of chi-square = 409.04 DF = 149 p = .000

Percent of variance explained by factor 1 = 22.53%
Percent of variance explained by factor 2 = 1.96%

Number of common factors = 3

Chi-square = 42979.69 DF = 284
Change of chi-square = 383.21 DF = 148 p = .000

Percent of variance explained by factor 1 = 25.39%
Percent of variance explained by factor 2 = 2.35%
Percent of variance explained by factor 3 = 1.42%

Subtest # 1 (Paragraph)

# of common factors = 2

Chi-square = 23592.89 DF = 538
Change of chi-square = 161.857 DF = 49 p = .000

Percent of variance explained by factor 1 = 34.34%
Percent of variance explained by factor 2 = 2.30%

Number of common factors = 3

Chi-square = 23508.16 DF = 490
Change of chi-square = 84.728 DF = 48 p = .001

Percent of variance explained by factor 1 = 33.01%
Percent of variance explained by factor 2 = 2.56%
Percent of variance explained by factor 3 = 1.24%

Subtest # 2 (Situation)

# of common factors = 2

Chi-square = 26337.88 DF = 531
Change of chi-square = 119.043 DF = 49 p = .000

Percent of variance explained by factor 1 = 19.59%
Percent of variance explained by factor 2 = 2.25%

Number of common factors = 3

Chi-square = 26234.20 DF = 483
Change of chi-square = 103.678 DF = 48 p = .000

Percent of variance explained by factor 1 = 19.25%
Percent of variance explained by factor 2 = 2.46%
Percent of variance explained by factor 3 = 2.12%

Subtest # 3 (Fill-the gap)

# of common factors = 2

Chi-square = 23958.79 DF = 507
Change of chi-square = 84.268 DF = 48 p = .000

Percent of variance explained by factor 1 = 23,28%
Percent of variance explained by factor 2 = 3.61%

# of common factors = 3

Chi-square = 23784.39 DF = 459
Change of chi-square = 174.40 DF = 48 p = .000

Percent of variance explained by factor 1 = 28.29%
Percent of variance explained by factor 2 = 2.26%
Percent of variance explained by factor 3 = 1.69%

Results of essential dimensionality approach

(AT1 items from Domain A)

Calculation summary for subtest # 1 (Paragraph)

AT lengths: m= 12
PT lengths: n = 26
# of partition cells: K = 16

Original examinee DIMTEST statistic group size: 694
Size after deleting sparse cells: 624

T_L = 5.591
T_B = 4.496
T = .775 p = .22

Calculation summary for subtest # 2 (Situation)

AT lengths: m= 12
PT lengths: n = 26
# of partition cells: K = 11

Original examinee DIMTEST statistic group size: 681
Size after deleting sparse cells: 618

T_L = 6.556
T_B = 3.908
T = 1.872 p = .031

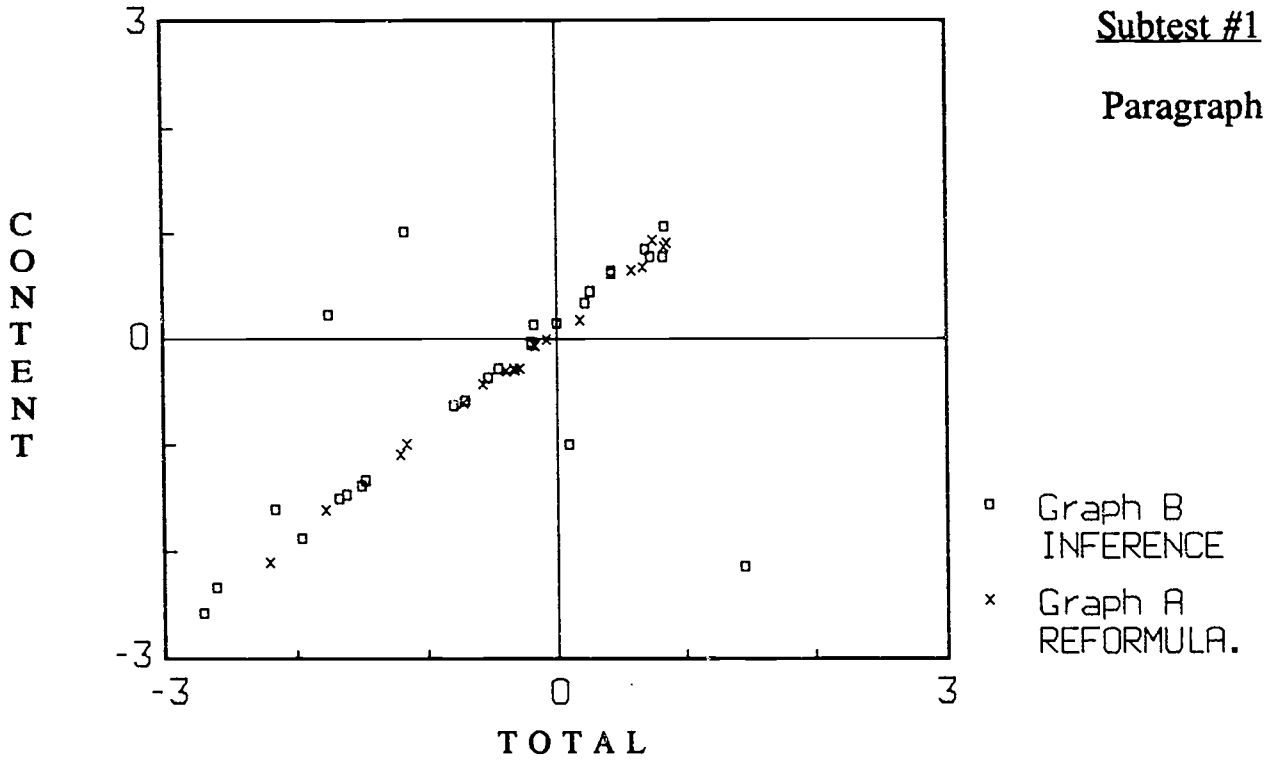
Calculation summary for subtest # 3 (Fill-the-gap)

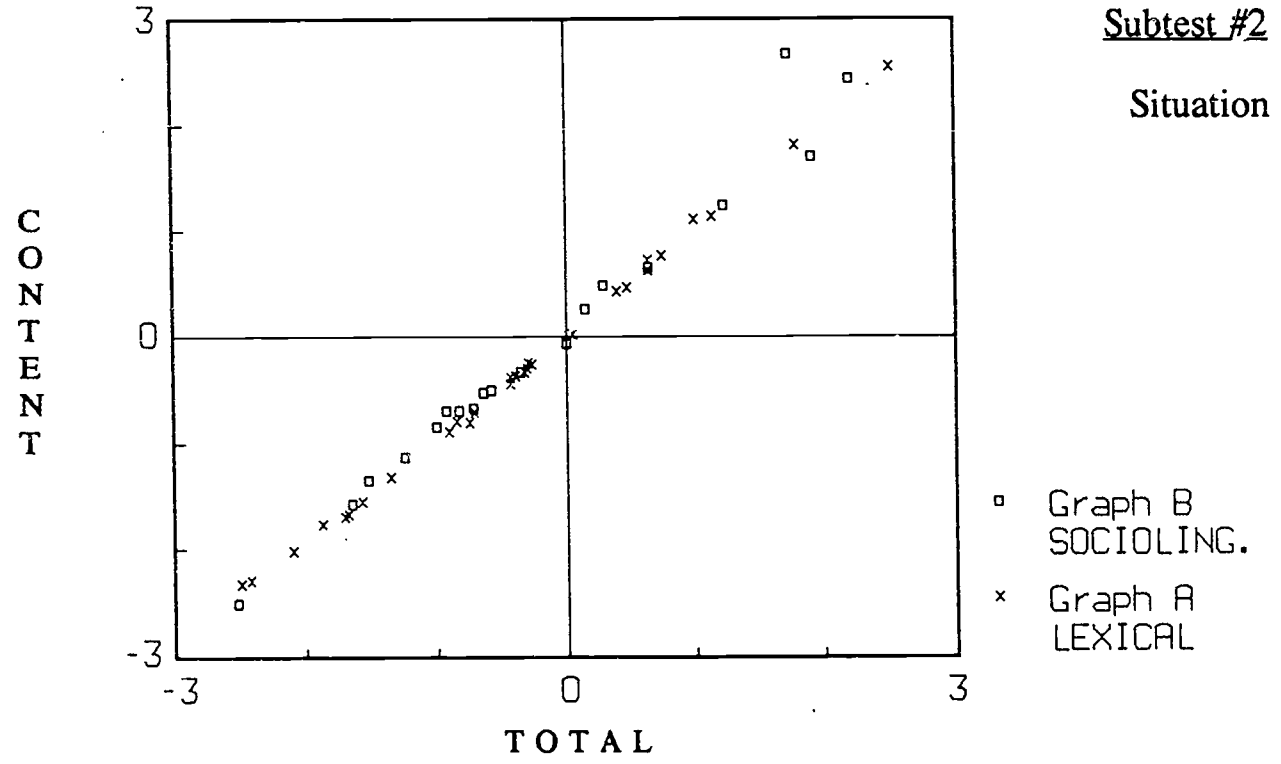
AT lengths: m= 12
PT lengths: n = 26
# of partition cells: K = 15

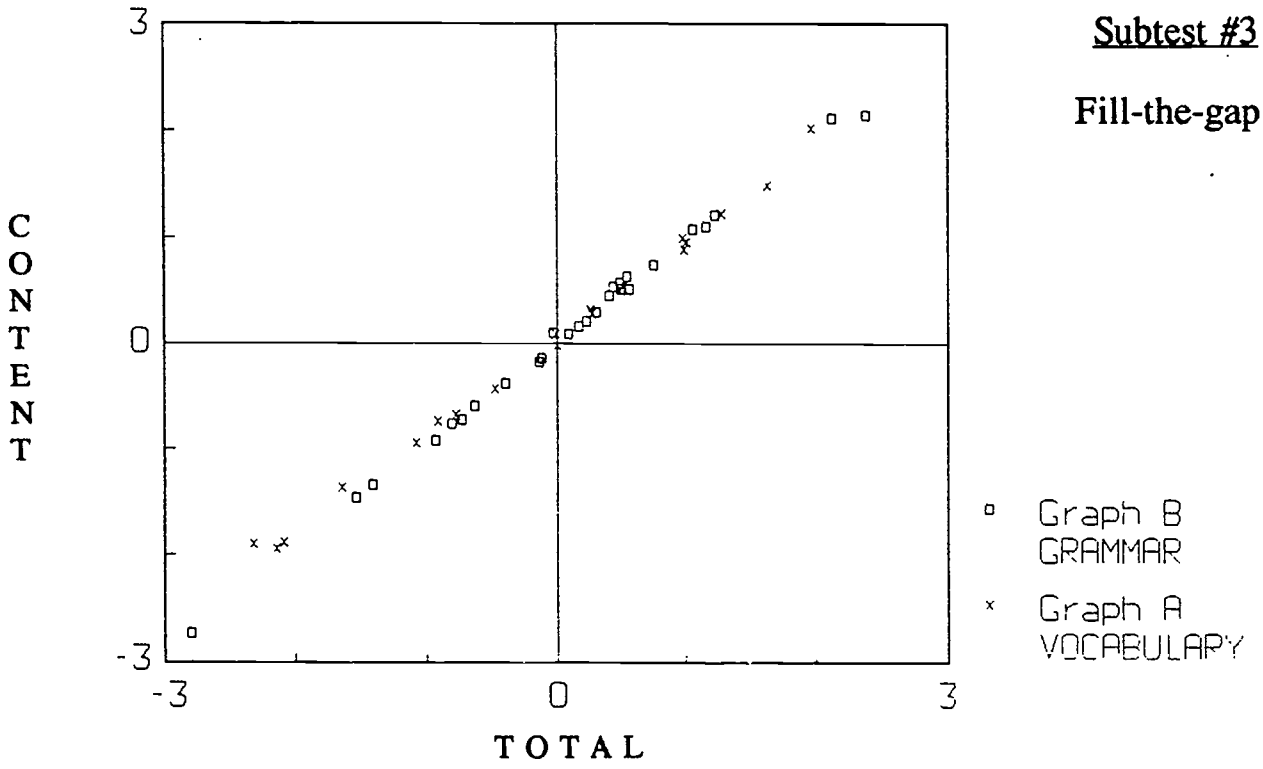
Original examinee DIMTEST statistic group size: 661
Size after deleting sparse cells: 573

T_L = 5.248
T_B = 2.833
T = 1.707 p = .044

Scattergrams of the Bejar's method







Subtest #3

Fill-the-gap