ED 364 063 FL 021 378

AUTHOR Des Brisay, Margaret; And Others

TITLE Opportunities for Input as Variables in the Trait Structure of Second Language Data Sets.

PUB DATE Aug 93

NOTE 24p.; Paper presented at the Annual Language Testing Research Colloquium (15th, Cambridge, England, and Arnhem, The Netherlands, August 2-8, 1993).

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Cloze Procedure; Comparative Analysis; *English (Second Language); Factor Analysis; Foreign Countries; French; Higher Education; Item Analysis; *Item Response Theory; *Language Role; *Language Tests; Native Speakers; *Second Languages; Statistical Analysis; *Student Experience; Test Items

ABSTRACT

This study investigated the appropriateness of Item Response Theory (IRT) models to analyze language test data. In a university ond language institute, results of typical versions of tests of Fr ch as a Second Language (n=831 examinees) and of English as a Second Language (n=617 examinees) were subjected to non-linear factor analysis in dimensionality assessment, as a first step in selecting an appropriate IRT model for data sets from students with differing language learning backgrounds. The tests, similar in design and purpose, each included a reading comprehension, listening comprehension, and cloze portion. Examinees were native speakers of one of the two languages tested, but were heterogeneous with respect to their language backgrounds (immersion programs, regular high school programs bilingual families, frequent contact with native speakers). No systematic relationship was found between item performance and strategies hypothesized to characterize each of the two examinee groups, but it is suggested that certain response patterns in the cloze portion of the test merit further investigation. (Author/MSE)

15th Annual Language Testing Research Colloquium
Cambridge/Arnhem August, 1993

Opportunities for Input as Variables in the Trait Structure of Second Language Data Sets

Margaret Des Brisay University of Ottawa
Lise Duquette University of Ottawa
Mohamed Dirir University of Massachussets

The unidimensionality of language test data has been a contentious issue for over 15 years with most studies reaffirming a unitary factor statistical model for language ability. Even when more than one factor was found, additional factors appeared to be neither construct nor skill-related (Davidson, 1988).

This paper reports the results of factor analyses undertaken to investigate the appropriateness of IRT models to language test data from typical versions of a test of Français langue seconde and a test of English as a Second Language. (N= 671 to 931). The tests are written to the same set of specifications and used for the same purpose.

Data were analyzed using both linear (SAS) and non linear factor analysis. The non-linear model used was NOHARM (Fraser, 1986). Non-linear analysis is preferred for dimensionality assessment as it does not assume linear relationships among the variables but gives loadings with are probabilistically related to the construct. While unidimensional models do fit subpopulations assumed to be more homogeneous with respect to learning experience, both linear and non-linear models suggest a two (ESL) or three (FLS) factor for entire data sets.

In most studies, examinees have had fairly uniform language learning opportunities (classroom centred) and low opportunity for input from native speakers. In the present study, examinees were native speakers of one or the other of the two languages being tested but were much more heterogeneous with respect to the process by which they had acquired L2; immersion programs, regular high school programs, bilingual families, frequent contact with native speakers. Difficulty in satisfying model assumptions and obtaining invariance of item parameters motivated efforts to relate the factorial structure of these test data to the language learning experiences of the examinees. Although no systematic relationship was found between item performance and strategies hypothesized to characterize each of the two groups of examinees, certain response patterns in the cloze section merit further investigation.

1. Introduction

Item Response Theory has been described as "undoubtedly the most striking development of the past several decades in educational measurement" (Carroll, 1990). An IRT model describes the relationship between the trait being measured and each of the items on the test by a mathematical expression (the item-characteristic function) that relates the probability of giving a correct response to the trait or ability being measured and the characteristic of the item. Where model assumptions can be met, IRT offers sample-free calibration facilitating item-banking, test-free person measurement facilitating equating of alternate versions and multiple reliability indices (information functions) facilitating the construction of tests which give optimal information at ability levels of interest.

The potential benefits of IRT will not be realized, however, unless invariant ability and item parameters can be estimated and invariant parameters will not be estimated if the data do not fit the assumptions of the chosen model. The first step, then, in investigating the utility of IRT models for any application must be the gathering of evidence to support the use of the model. The one-, two-, and three-parameter models commonly used (eg.,BILOG, LOGIST,) share the assumption of unidimensionality; that is, the assumption that only a single ability is necessary to account for examinee test performance. There are certain *a priori* grounds for believing the unidimensionality assumption a difficult one to meet in the case of language tests. Psycholinguistic models of language proficiency as proposed, for example, by Canale and Swain (1980), and Bachman (1991) allow for different examinees to use different composites of skills in producing a correct response. Swain (March,1993, personal communication) expresses doubt that any *good* test of communicative competence can satisfy the assumptions necessary for IRT.[1] But examinees and items interact and as Ackerman (1992) stresses it is, in fact, this interaction which must be unidimensional. No matter what variability exists among the examinees, if all the items are measuring a single skill, the interaction will be unidimensional. If the items measure several skills but the examinees vary on only one of these or on the same composite of skills, again the interaction will be unidimensional.

2. Background to Present Study

It is frequently observed that the dimensionality of a test can vary from one group of examinees to another (Ackerman, 1992). Earlier research at the University of Ottawa, Second Language Institute, (Hambleton et al, 1992, Des Brisay and Laurier, 1990) with data from the Canadian Test of English for Scholars and Trainees (CanTEST) had provided clear evidence supporting the use of IRT (the two parameter model). Both linear and non-linear factor analysis supported the assumption of unidimensionalty, the Bejar (1980)[2] confirmed the invariance of item parameters, and the accuracy of model predictions, assessed through an analysis of BILOG residuals, indicated good model data fit. However, almost all CanTEST examinees are tested

[1] Multidimensional IRT models exist but technical developments have been limited to date and very large samples are required to obtain satisfactory parameter estimates.

[2] The Bejar method involves identifying subsets of items for which there are *a priori* grounds for hypothesizing an additional dimension and estimating item parameters for these items, both separately and as part of the test total. If the data are unidimensional, the plot of the subset- and total test- based difficulty parameters should be close to a line or theoretical axis with a slope of 1.0 and an intercept of 0.0.

overseas in Beijing or Jakarta, are male, age 24 to 40, university educated (engineering or agriculture) and have very similar classroom language learning experiences.

The Second Language Institute is involved in the production of two other second language proficiency tests, one a test of ESL and the other a test of français langue seconde (FLS). These tests are administered five times a year to undergraduates at the University of Ottawa to determine whether or not they meet the level of SL proficiency which is a degree requirement at this university. As is common practice today, the Second Language Proficiency tests are compiled from banks of pre-trialled sets of passage related items. This avoids the security problems associated with reusing an intact version and demands less labour than would be involved in producing and equating an entirely new version. Test construction decisions are based on number of factors - desired content coverage, pedagogical impact, administrative constraints, and security concerns. First, however, the items in the bank must be calibrated so as to ensure that different versions have comparable statistical properties with respect to level of difficulty, discrimination and precision of scores.

In September, the clientele for the ESL/FLS Proficiency tests is largely composed of newly admitted students; in December, there is a large percentage of students who failed to meet the requirement in September and in April, an even larger number of repeaters. This means that items will almost certainly have different difficulty indices at different administrations. IRT offers a way to dealing with this problem by placing test item statistics obtained from non-equivalent samples of examinees on a common scale. Experience gained applying an IRT model to the CanTEST motivated an investigation into the utility of the same measurement models for the ESL/FLS testing program with its much more heterogeneous test clientele.

3. The ESL/FLS Proficiency Tests

The ESL/FLS Proficiency tests are designed to measure general comprehension rather than formal knowledge of the language. University regulations permit students to submit written work in their first language and faculty must have an SL proficiency that allows student-teacher consultations to be conducted in the student's first language; consequently, there is no need to measure productive skills where accuracy would be of more concern.

New combinations of passages and related items are compiled for each of the five annual administrations but the descriptive statistics shown in **Table 1** are typical for the September population. The tests are entirely multiple choice. Each test has three parts; listening comprehension (12 to 15 items), reading comprehension (12 to 15 items) and a cloze passage with 30-32 blanks. The sections are weighted to count for one third of the total mark each and decisions are based on total score.

Insert Table 1 about here

These data are typical in that the cloze passage is the easiest subtest for Anglophone students writing the French test and the most difficult subtest for Francophone students writing the English test. Possibly as a consequence, the cloze passage is typically less reliable for the French test population. For the overseas (CanTEST) populations, the range of scores for the cloze is much narrower (sd=4.2 vs 7.2) and again, the cloze subtest is both less reliable and less highly correlated with other test components. However, the descriptive statistics for all three tests are very similar.

Care is taken in the elaboration of listening and reading comprehension items to focus on the general meaning and organization of texts, to limit items focusing on specific details and to avoid items that could be correctly answered from grammatical information alone. The tests are designed to reflect and encourage real life language use and, as a result, could be considered to advantage examinees who have had out-of-class learning opportunities. (On the other hand, it could be argued that any test will advantage the book-learner.)

The texts chosen for the cloze passages, like those for other sections of the tests, are authentic documents intended for native speakers taken from journals, magazines, textbooks, introductory lectures, chosen in function of the background knowledge and educational level of the examinees. Given that the emphasis is on measuring overall comprehension and discourse processing, departures from the principle of random deletion occur in the construction of the clozes to ensure that the majority of blanks focus on lexical rather than grammatical elements of the passage. This seems to be particularly necessary in construction of the French cloze passages (textes lacunaires) where random deletion will lead to a large number of items involving determiners.

4. FLS/ESL Examinee Characteristics

Canadian students who write the ESL/FLS Proficiency tests at the University of Ottawa proficiency tests have acquired their second language in a variety of educational, social or professional settings. All students will have had some formal instruction in their second langauge but education in Canada is a provincial responsibility, and both the content and methods of evaluation in second language programs vary considerably from province to province. Certain students are the products of immersion programs (early or late) and are very advanced in their second language particularly with respect to oral interaction. Grammmar and written production may not have been mastered, however. Those from regular SL programs may have studied a minimum of three years or a maximum of 13 years, again depending on the province and even on the resources of the individual school boards.

Examinees will also vary with respect to their out-of-class SL experience; some come from bilingual parts of the country, (New Brunswick, the Ottawa area), some from regions which are predominantly Anglophone (Alberta) or predominantly Francophone (Saguenay-Lac St. Jean); some have had work experience in their second language and some even come from bilingual families. (It is not rare to have students who are unsure of which language to designate as their mother tongue).

Those who achieve over 50% are exempt from compulsory second language courses but may register for advanced level courses or may even major in their second language (French only). The heterogeneity of the test clientele is evidenced in the make-up of the language classes at the Second Language Institute. Francophone students are typically observed to be stronger in listening and speaking than they are in reading and writing. This is hardly surprising given the Canadian reality. Francophones, especially Franco-ontarians, will invariably have had more oppoitunities to acquire their SL in a natural setting. Such opportunities are available less often to Anglophone Canadians and, of course, they are not available at all to examinees at overseas test sites. Examinees in an EFL context with low or no opportunity for input tend to rely more heavily on their knowledge of the formal aspects of the language in performing any language task. It seems reasonable to assume that these different SLA opportunities, some learning intensive and some acquisition intensive have resulted in different configurations of language proficiency. Indeed, it is commonly observed that students with similar scores on any proficiency test will have different strengths and weaknesses. In particular, teachers at the Second Language Institute find that cloze tasks give a strong indication of which students "know their grammar" and which " use effective communication strategies". (Many students, of course, display both types of knowledge but they are usually exempt from language classes.) Teachers also find that these differences frequently coincide with what is known about the context in which the students' SL was learned/acquired; that is, whether there was high opportunity or low opportunity for input. Further variation will exist within groups in function of other factors (motivation, aptitude, intelligence) which are not determined by learning/acquisition opportunities but which mediate their impact.

The Present Study

1. Purpose

The primary purpose of this study was a methodological one: to investigate the utility of non-linear factor analysis (FA) in testing the IRT model assumption of unidimensionality as a first step in selecting an appropriate IRT model for the FLS/ESL Proficiency Tests. The two data sets analysed were ones in which there were *a priori* reasons (high versus low opportunities for input) for expecting violations of this model assumption.

Given that it is poosible for a model to have applied utility even if model assumptions are being violated, two other types of evidence supporting model use as recommended by Hambleton and Swaminathan (1985) were examined: the extent to which the expected properties of the model (ie, invariance of item and ability parameters) were obtained, and the accuracy of model predictions using real test data.

The study was also extended to assess the interpretability of the factor loadings obtained from the non-linear FA used in an exploratory manner. It was hypothesized that differences in examinees as related to their language learning situation (high versus low opportunity for input) would be reflected in the factor loadings for cloze items. A final interest was to compare both the dimensionality and factor structure of the two data sets to determine their impact on model choice. Are the same models valid for both French and English data sets?

2. Procedures

The procedure selected for the study of dimensionality was normal harmonic factor analysis (Fraser & McDonald, 1988) as implemented in the **NOHARM** program. For purposes of comparison **SAS** was also used to perform linear factor analysis. Both the Bejar (1980) method (not reported on in this paper) and an analysis of residuals using the program **RESID** were used to further investigate model-data fit. The IRT computer program **BILOG** (Mislevy & Bock, 1986) was used to obtain item parameter estimates to use as input to both these procedures. The data were analyzed were from the September test administrations for 1991 and 1992 of the ESL/FLS Proficiency tests. Findings were similar and only 1992 data are reported in Tables 1-9.

Linear factor analysis is the most widely used method in dimensionality assessment. The standard procedure with binary items is to obtain the tetrachoric correlations among the items, get the principal components or common factors and examine the eigenvalues of the correlation matrix. The magnitude of the eigenvalues, the differences between successive eigenvalues or the amount of variance explained by the factors and an inspection of factor loadings are all pieces of evidence to support hypotheses regarding the dimensionality of the data set. However, Hattie (1985) argues that linear factor analysis is not the perfect choice for assessment of the dimensional structure of binary data as a linear relationship among the variables and the factors cannot be assumed. Further support for the use of non-linear factor analysis is found in Takane & De Leeuw (1987), cited in Gessaroli (1991) where it is shown that the model used in IRT and non-linear factor analysis are mathematically equivalent. The use of non-linear factor analysis in dimensionality assessment takes advantage of this relationship.

The nonlinear factor analysis program, **NOHARM**, which was used in this study fits unidimensional and multidimensional normal ogive models using a least squares procedure which seeks to minimize the squared differences between the observed sample and the estimated bivariate proportions correct. Although NOHARM uses linear approximations, it is non-linear in both coefficients (parameters) and latent traits. According to McDonald (1981), the size of the NOHARM residuals in a measure of the departure from independence. Carlson & Jirele (1992) have reported a successful application of non-linear factor analysis in assessing dimensionality. Gessaroli (1991) found non-linear factor analysis to show a fairly high rejection rate of unidimensionality when two-dimensional data were generated and an incremental fit index used.

3. Additional Tests of Model-Data Fit

The major assessment of goodness of fit in this study involved an analysis of the residuals. This was done using the computer program **RESID.** RESID accepts either BILOG or LOGIST output as input and computes standardized residuals based on the difference between observed and expected (ie, predicted by the model) performance on individual items for a specified number of ability groups. The default number of 12 was used with these data. Residuals are recommended (Bachman, 1992, Hambleton et al, 1992) for assessing goodness of fit. Residuals should be small and random and standardized residuals (residuals with the error term removed) should be normally distributed (mean = 0, standard devia -- 1) when model fit is good.

4. Exploratory Factor Analysis (Trait Structure)

Factor loadings obtained from both linear and non-linear analysis for the 2-factor solution (ESL) and the 3-factor solution (FLS) were also examined. Studies into the factor structure of language tests (eg., Davidson, 1988) using linear techniques, have consistently failed to produce interpretable factor loadings. One problem has been that, even when tetrechoric correlations are used, the second factor often appears to be difficulty. As McDonald (1981) views it, the problem is not difficulty but linearity. McDonald reports that a major consequence of using linear factor analysis on binary items is to distort the loadings of the very easy and very difficult items and to make it appear that such items do not measure the same underlying dimensions as the other items.

5. Units of Analysis

Non-linear factor analysis was performed on test total, on cloze passages alone and on High (65% or above) and Low (below 65%) scorers. Initially we had intended to use demographic data in order to form subgroups but this was not possible. The decision to divide the sample at 65% was based on the collective experience of language teachers at the Second Language Institute and also on study conducted by the Ottawa (Secondary) School Board on FLS proficiency test performance of graduates from the different board programs. Both sources indicated that it would be rare for an examinee who had not had high opportunity for input to obtain above 65%. However, without confirming demographic data, the division can only be interpreted as separating high and low test performers.

6. Content Analysis of Cloze Passages

The cloze passages are considered the test component in which items tapping grammatical knowledge and those tapping comprehension of overall meaning could be most easily identified. It was for this reason that the cloze passages were subjected to NOHARM analysis both separately and together with the other items on the test. It should be noted that the nature of cloze passages may lead to violations of the principle of local independence. We return to this point in the discussion.

Items in the cloze passages, particularly those on the FLS test, had been classified into content and function words in the test construction phase. Post-hoc content analysis was conducted following the results of the exploratory non-linear factor analysis to further investigate the relationship between the factor loadings and the generic properties of the items. In the post-hoc analysis, cloze test items were further classified according to whether a correct response involved attending to meaning *and* form, attending to form only or attending to meaning only. Most items are designed so that all options are grammatically correct. However, some items have distractors which can be eliminated on grammatical criteria and a successful test-taking strategy might involve using grammatical knowledge to arrive at the right answer by the process of elimination. Items involving cognates were also identified. French and English share a large number of cognates and students are commonly observed to vary in their ability to exploit this. Duquette (1993) found that the most common strategy of intermediate students, who had not received any training in strategies for guessing meaning from context, was to look for roots that were common to French and English (cognates and borrowings) whereas more advanced students are able to exploit the external clues found in the sentence or the passage.

6. Results of Dimensionality Assessment

As our main purpose was to the investigate the usefulness of non-linear factor analysis and, in particular, the program NOHARM, in the assessment of unidimensionality , our initial interest was not in the loadings per se but in the degree to which a unidimensional model could account for the data.

Fraser (1986) suggests that a root mean square residual (RMSR) < 4/sqrtN be taken as an indicator of the goodness of fit of a one dimension (factor) solution and all tests (total or components) do meet this criterion. As can be seen in **Table 2**, RMSR for a one dimension NOHARM solution ranged from .0212 for the ESL Low cloze response data to .0122 for the FLS High cloze response data. Better fit (ie., smaller residuals) is obtained in all cases with the addition of a second factor and RSMR is higher for low performers than for high performers.

insert table 2 about here

Although Fraser describes the RMSR as only a "rough indication" of model fit, he expresses doubt that a more refined test of significance would reject the hypothesized model. However, Gessaroli (March 1993, personal communication) has found in Monte Carlo studies that this test fails to detect the multidimensionality of simulated data and recommends instead an index based on the distribution of standardized residuals. According to Gessaroli's index (PERZ), 5% or less of the standardized residuals in the variance-covariance matrix would be expected to exceed ± 1.96 after fitting a one-factor solution to unidimensional data under well-fitting model conditions. Both of the ESL and FLS proficiency tests fail to meet this criterion after fitting a one dimensional solution. The FLS test, furthermore, fails to meet the criterion after fitting a 2 dimensional solution. Findings are similar when the cloze items are analyzed separately. Dividing the sample into high and low performers does not seem to affect the PERZ index for the FLS test, but 12.4% of residual covariances exceed the PERZ index of ±1.96 when responses for the high performing examinees on the ESL test as compared with 21.4% for responses from the low performing examinees. Table 3 shows PERZ for test totals and the cloze passages for the entire sample and for cloze passages with high scoring (65 % and over) and low-scoring (under 65%) examinees.

insert Table 3 about here

This contrasts with the NOHARM analysis of CanTEST data which consistently produces fewer than 5% of residuals > ± 1.96 when a one dimension solution is requested. Earlier research by Ready (1991) with low level Anglophones on the reading comprehension and cloze sections of a French placement test used at the University of Ottawa had also found a one dimension solution using NOHARM provided the best fit with PERZ > ± 1.96 = 3.6%.

The eigenvalues and amount of explained variance from linear factor analysis also suggest a three factor solution for the FLS data and a two factor solution for the ESL data (see Table 4 and 5). In both cases , considerably less that 20% of the variance is explained by the first factor.

insert tables 4 and 5 about here

Analysis of Residuals (Accuracy of Model Predictions)

The results of the residual analysis for the 1-, 2-, and 3-parameter models for the data sets are shown in **Tables 6 and 7** and strongly suggest that an appropriate IRT model could have applied utility for these data.

---

insert Tables 6 and 7 about here

Two additional tests of fit are also shown in **Tables 6 and 7.** Both are chi-square fit statistics, one computed by BILOG and the other, Q_1, (Yen, 1981) computed by RESID. Both these statistics have been shown to sensitive to sample size in simulations studies. In one such study (Hambleton et al, 1992) the number of items flagged as misfits (Q_1) increased from 1 with a sample size of 150 to 6 with a sample size of 600.

In the case of the English and French Proficiency tests, the two parameter model would appear to be appropriate for operational use. The standardized residuals for both data sets are considerably smaller for the two- and three-parameter models than they are for the one-parameter model. The percentage of biserials exceeding the mean biserial ± .15 was 23% for the FLS data and only 3% for the ESL data. This suggests that the assumption of equal discrimination is tenable only for the ESL data . The residual analysis gives no compelling reason to choose the three parameter model, even in the case of the French test which is difficult for its population and where there is known to be considerable guessing.

8. Interpretation of Factor Loadings

As McDonald's finding with respect to very easy and very difficulty items implies, results using linear models of tetrachoric correlations are more likely to be similar to results obtained from non-linear FA when all items are of moderate difficulty. This proved to be the case with the data used in this study. Patterns of factor loadings were obtained with both linear and non-linear analysis (see **Tables 8 and 9**). As an unexplained result of the rotation procedures used, however, factors 1 and 2 are reversed.

insert Tables 8 and 9 about here

All attempts to relate the factor loadings of groups of items to generic properties of those items were unsuccessful. Loadings appeared to be neither skill nor passage related. There were no systematic relationships between groups of items and their loadings with respect to classical indices for difficulty and discrimination. Post-hoc inspection of items suggested no pattern related to proximity of clues to the blank, the nature of the clues (semantic or grammatical), or the presence of cognates.

As was seen in **Table 5**, somewhat larger residuals are associated with cloze test items, a possible indication of departure from independence. Individual items with residuals larger than .05 were examined but no pattern was detected. (Nor did an examination of tetrachoric correlations (PRELIS) suggest violations of the principle of local independence.)

One curious pattern in the factor loadings for cloze items was noted. There was a tendency for two items in the same sentence to load on different factors. This would suggest that two items within a sentence behave more as independent observations than do two 'first' items from different sentences, hardly consistent with either departures from independence or the exploitation of contextual clues and overall meaning as successful cloze-taking strategies.

Discussion and Conclusions

We have already emphasized that the factor analysis reported on in this study was not undertaken in the interests of validating a theoretical construct pertaining to language proficiency or language acquisition. Indeed, none of the instruments discussed had been constructed to measure hypothesized dimensions of language proficiency. Rather, the main purpose of the study was to investigate the use of non-linear factor analysis in dimensionality assessment as a first step in selecting an appropriate IRT model for data sets obtained from examinees with differing learning/acquisition backgrounds.

All modelling requires some suspension of disbelief. A measurement model must be simpler that the mental construct being measured or it is not much use as a model. This should be kept in mind when testing model assumptions such as that of unidimensionality. According to Ackerman (1992), any test of more than one item is never exactly unidimensional. Hattie (1985), in a review of existing indices for dimensionality assessment, suggests "it is probably more meaningful to ask the degree to which a set of items departs from unidimensionality than to ask whether a set of items *is* unidimensional."

Although departures from unidimensionality were detected, no clear relationship between violations of unidimensionality and lack of model data fit could be established. In spite of the different results from checks on dimensionality and invariance, the analyses of residuals are very similar for both tests and are similar as well to those typically found for CanTEST from overseas testing sites with their more homogenous test clientele. **PERZ** may be too rigorous a test of significance and researchers should consider using **RMSR** as recommended by Fraser in deciding whether or not a unidimensional model gives a satisfactory account of the data. The lack of model data fit for the one-parameter model for the ESL/FLS data sets may even suggest that other model assumptions such as that of equal discrimination of items are more critical in fitting IRT models to test data.

Certainly in a testing system where language proficiency is reported with a single score, unidimensionality is a useful concept. Indeed, it is hard to see how the score can be interpreted unless the underlying trait is assumed to be one that can be modeled unidimensionally. At the same time, it is recognized that "a score reflects a complex combination of processing skills, strategies and knowledge components, both procedural (process) and declarative (content), some of which are invariant and some variant across persons, tasks or stages of practice" (Snow and Lohman, 1989). If the components of this complex combination are all valid in terms of language proficiency then the single score can be justified as a summary statistic for many purposes. In interpreting scores, however, one needs to know if group membership is characterized by position on an underlying trait, albeit a complex one, or whether group membership is based on some nuisance variable such as sex, mother tongue, or background knowledge.

In this study we explored the impact on test dimensionality of group membership based on differing opportunities for input (learning intensive or acquisition intensive). The examination of factor loadings was conducted to find a possible explanation for the seeming failure of the data to meet the model assumption of unidimensionality. Although the data are more factorially complex than data from overseas testing sites, no causal link between departures from unidimensionality and observed differences in the knowledge base of examinees could be made. Nor were we able to substantiate any of the hypothesized relationships between factor loadings and item content.

Ceıtain studies (Cziko (1978), Maclean and Anglejan (1986) reveal that L2 students have difficulty exploiting contextual constraints that operate above the sentence level and that it is only at the advanced level (or among native speakers) that semantic and discourse constraints are effectively used. Other studies in L1, (Sternberg et al., 1982, 1983) show vocabulary to be the better predictor of overall comprehension. Studies in L2 (Marton (1977), Meara (1980) and Richards (1985), show vocabulary to pose difficulties even at the advanced level . But at the University of Ottawa, the analysis of cloze test results (FLS) have revealed that intermediate and advanced students have more difficulty with the grammatical elements of the cloze than with the semantic ones. Whether the difference between second language and foreign language settings offers a partial explanation of the trait structure of these data remains a subject for further investigation.

In future research other models, for example, cluster analysis, will be considered. Data for future research will include data from a study to validate a model of acculturation developed by Clément & Noels (1992). In addition to writing cloze tests taken from the ESL/FLS proficiency test bank, subjects in the Clément & Noels study completed questionnaires detailing the quantity and quality of their SL contact. Data were also obtained from Statistics Canada confirming the minority/majority status of subjects' mother tongue in the locations where they reported living longest. Access to these data must be approved by the University of Ottawa's Ethics Committee.

The possibility that the nature of cloze passages may lead to violations of the principle of local independence also deserves further investigation. However, since items in a cloze passage are always presented in the same order, violations of this principle are not so serious as they would be in a case where the items involved would be re-used in different or unpredictable contexts, as in computer adaptive testing, for example. However, if the principle of local independence is being violated, that is, if success on one item in a cloze passage is influenced by success or failure on a nearby items or items, there will be less information in the subtest than the sum of the item information functions would indicate. (Moreover, in applications of classical measurement theory, reliability estimates will be spuriously high). This must be taken into consideration for such IRT applications as optimal test design where a target test information function must be specified. Technical reservations about the cloze subtest must be set against the conviction of the teachers at the Second Language Institute, French and English, that the cloze passages do a good job of "sorting students out".

The major concern of the test constructor remains making the best estimates of ability possible and ensuring that such estimates are being made fairly and consistently across different administrations and versions. Test developers considering the adoption of an IRT model must be prepared to conduct a thorough investigation into model appropriacy and utility with their own data. In the case of the ESL/FLS Proficiency tests the issue was whether or not the heterogeneity of our test populations had consequences for our choice of a measurement model to guide decisions about test construction and examinee performance. The data may be more factorially complex than would be the case with a homogeneous populations but the accuracy of model predictions as evidenced by the analysis of BILOG residuals indicates that a two parameter unidimensional IRT model could have applied utility for these tests. However, comparison of the results of this study with those from earlier studies with CanTEST item response data would not support the use of item calibrations obtained from responses of overseas examinees in an EFL environment with those obtained from Canadian students being tested in their second language.

References

Ackerman, T. (1992). A Didactic Explanation of Item Bias, Item Impact and Item Validity from a Multidimensional Perspective. *Journal of Educational Measurement, Vol. 29, No.1.* pp.67-91

Bejar, I. (1980). A procedure for investigating the dimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17.*

Bachman, L.F. (1990) *Fundamental Considerations in Language Testing*, Oxford. Oxford University Press

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

Carlson, J.E., & T. Jirele (1992). Dimensionality of 1990 NAEP Mathematics Data. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1992.

Carroll, J.B. (1990). Future developments in educational measurement in Walberg, H.J. and G. D. Haertel, eds., *The International Encyclopedia of Educational Evaluation*. Oxford. Pergamon Press.

Choi, Inn-Chull and Bachman, Lyle F. (1992) An investigation into the adequacy of three IRT models for data from two EFL reading tests. In *Language Testing, Vol 6.*

Clément, R., & K.A. Noels. (1992) Towards a Situated Approach to Ethnolinguistic Identity: The Effect of Status on Individuals and Groups. *Journal of Language and Social Psychology. Vol 11, No 4.*

Cziko, G.A. (1978). Differences in first and second language reading; the use of syntactic, semantic and discourse constraints. *The canadian Modern Language Review, Vol 34. No. 3.* 473-489.

Davidson, F. G. (1988). An Exploratory Modeling Survey of the Trait Structure of Some Existing Language Test Datasets; unpublished Ph.D dissertation. University of California at Los Angeles.

Des Brisay, M & M. Laurier (1991). Developing Small Scale Standardized Tests Using an Integrated Approach. *Bulletin of The Canadian Association of Applied Linguistics,* Vol 13, No. 1. 57-72

Duquette, L. (1993). L'étude d'apprentissage du vocabulaire en contexte par l'écoute d'un dialogue scénarisé en français langue seconde. Publication B-187. Québec. Université Laval, CIRAL.

Fraser, C., & R.P. McDonald (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23,* 267 -269

Fraser C. (1988). NOHARM: *A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* Armisdale, Australia. The University of New England, Centre for Behavioral Studies.

Gessaroli, M.E. (1991). Assessing Test Dimensionality Using an Index Based on Non-Linear Factor Analysis. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. April 1991.

Hambleton, R.K., Des Brisay M., & M. Dirir (1993). New Measurement Models and Methods for Constructing Language Tests. *Carleton Papers in Applied Language Studies*, Vol 10. Carleton University Press.

Hambleton, R.K. and Swaminathan, H. (1985). *Item response theory; Principles and Applications.* Boston. Kluwer.

Hattie, J. (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement, Vol. 9.* pp 169-164.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34,* 100-117

Marton, W. (1977). Foreign vocabulary learning as problem No. 1 of language teaching at the advanced level. *Interlingual Studies Bulletin 2 (1).* 33-57.

McDonald, R.P. (1982) Linear versus nonlinear models in item response theory. *Applied Psychological Measurement,* Vol.6, No. 4. 379-396.

McLean, M.& A. d'Anglejan (1986). Rational cloze and retrospection; insights into first and second language reading comprehension, *The Canadian Modern Language Review., 42, No.4,* 814-826.

Meara, P. (1980). Vocabulary Acquisition: A Neglected Aspect of Language Learning, *Language Teaching and Linguistic Abstracts, 13,* 221-246.

Mislevy, R., & Bock, R.D. (1986) *BILOG: Maximum likelihood item and analysis and test scoring with logistic models.* Mooresville, IN: Scientific Software.

Richard, J. C. (1985) Lexical Knowledge and the Teaching of Vocabulary. In J.C. Richards, (ed), *The Context of Language Teaching,* Cambridge, Cambridge University Press.

Snow, R. E. & Lohman, D. F. (1989) Implications of Cognitive Psychology for Educational Measurement. In R. L. Linn, ed., *Educational Measurement,* 3rd Edition. New York. ACE/Macmillan.

Sternberg. R.J., & J.S. Powell (1983). Comprehending verbal comprehension. *American Psychologist,* 38(7), 878-893.

Table 1 Descriptive Statistics for Typical Data Sets

		N	K	X	Mean %	S.D.	SEM
ESL	Total	617	71	.92	62.1	13.3	3.7
	Cloze	617	32	.89	57.3	7.4	2.5
FLS	Total	831	56	.89	58.5	10.1	3.3
	Cloze	831	30	.83	66.0	5.7	2.4
CanTEST	Total	294	115	.91	68.4	15.9	4.8
	Cloze	294	30	.73	60.1	4.5	2.3

Table 2 NOHARM Root Mean Square Residuals

Data Set	Number of Factors	RMSR
ESL Total Test (N=617)	1	.0155
	2	.0091
	3	.0075
ESL Cloze (N=617)	1	.0165
	2	.0080
	3	.0072
ESL Cloze High N=314	1	.0146
	2	.0093
	3	.0083
ESL Cloze Low (N=303)	1	.0212
	2	.0112
	3	.0099
FLS Total test (N=831)	1	.0139
	2	.0101
	3	.0069
FLS Cloze All (N=831)	1	.0148
	2	.0114
	3	.0082
FLS Cloze High (N=375)	1	.0122
	2	.0089
	3	.0058
FLS Cloze Low (N=456)	1	.0185
	2	.0147
	3	.0137

Table 3 **Percent of Standardized Residuals > ± 1.96* (PERZ)**

		1 Dim	2 Dim	3 Dim
ESL	Total Test	20.8	4.4	2.0
	All Cloze		—	—
	High Cloze	12.4	3.4	1.4
	Low Cloze	21.1	1.3	1.1
FLS	Total Test	21.4	11.9	1.9
	All Cloze	—	4.4	—
	High Cloze	17.5	8.5	3.2
	Low Cloze	19.3	11.0	0.2
CanTEST	Total	3.9	2.3	—
	Cloze	—	—	—

* fewer than 5% indicates good model data fit

Table 4 Eigenvalues and Explained Variance for ESL Data

No. of Factors	**1**	**2**	**3**	**4**	**5**
Eigenvalues	11.06	3.91	1.73	0.8	0.69
Variance	15.6	5.5	2.4	1.1	1.0

Table 5 Eigenvalues and Explained Variance for FLS Data

No. of Factors	**1**	**2**	**3**	**4**	**5**
Eigenvalues	7.77	2.42	1.89	0.65	0.58
Variance	13.9	4.3	3.4	1.2	1.0

Table 6 **Analysis of Standardized Residuals English Proficiency Test**

	1P	2P	3P
< -3	1.08	0.27	0.67
-3 to -2	4.30	1.75	2.02
-2 to -1	15.05	10.75	11.83
-1 to 0	30.78	38.71	32.66
0 to 1	32.39	35.35	42.47
1 to 2	12.10	11.56	9.81
2 to 3	3.49	1.34	0.40
> 3	0.81	0.10	0.13
Q1 (p = .05)	16	5	7
AAR*	0.928	0.743	0.713
BILOG (p = .01)	12	0	0

* Average of absolute-valued standardized residuals; Normal = .790

Table 7 **Analysis of Standardized Residuals Test de compétence (FLS)**

	1P	2P	3P
< -3	1.64	0.30	0.30
-3 to -2	5.95	1.93	1.64
-2 to -1	15.77	13.84	13.54
-1 to 0	23.66	35.71	34.97
0 to 1	30.65	32.74	37.50
1 to 2	15.48	13.54	9.97
2 to 3	5.06	1.79	2.08
> 3	1.79	0.15	0.00
AAR*	1.102	0.82	0.75
Q1 (p = .05)	22	6	7
BILOG (p =) .01	19	1	1

Table 8 **Factor Loadings (Rotated) for ESL Cloze Items**

	NOHARM		SAS	
Item	Factor 1	Factor 2	Factor 1	Factor 2
1	.394	.397	.30933	.30878
2	.093	.709	.57034	.07121
3	.544	.204	.15747	.42138
4	.364	.205	.16294	.28227
5	.539	.016	.01474	.41573
6	-.281	.806	.65263	-.22839
7	.392	.056	.04469	.29033
8	.690	.059	-.05112	.55494
9	.082	.727	.58668	.06068
10	.115	.473	.36514	.08616
11	.696	.076	.05359	.55564
12	.137	.775	.62571	.09996
13	.529	.285	.21857	.40635
14	.577	.006	.01208	.45685
15	.028	.622	.49407	.02140
16	.604	.137	.11105	.48154
17	.243	.543	.43544	.19109
18	.263	.386	.31327	.20567
19	.528	.165	.13057	.42214
20	.205	.567	.46135	.15524
21	.508	.258	.20404	.40687
22	.464	.512	.41649	.36311
23	.320	.638	.52111	.25213
24	.627	.188	.15223	.50157
25	.381	.486	.38809	.29632
26	.558	.417	.33244	.43891
27	.323	.635	.51596	.24786
28	.410	.540	.43781	.32461
29	.527	.446	.35875	.41619
30	.457	.567	.46151	.36090
31	.455	.582	.47197	.36147
32	.535	.373	.30251	.42261

Table 9 Factor Loadings (Rotated) for FLS Cloze Items

	NOHARM			SAS		
Item	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
1	.202	.353	.304	.27859	.16028	.23742
2	.497	.352	.275	.27685	.40157	.22160
3	.239	.532	.464	.42816	.18355	.36407
4	.530	.397	.117	.31328	.42244	.08794
5	.291	.094	.273	.07864	.22190	.20971
6	.200	.239	.277	.18576	.16355	.21783
7	.002	.579	.203	.45167	-.00194	.15009
8	.531	.213	-.093	.16459	.40951	-.07030
9	.095	.170	.552	.12962	.07550	.42011
10	.079	.622	.018	.47034	.05924	.01585
11	.728	.206	-.166	-.15666	.57483	-.12799
12	-.238	.022	.777	.01367	-.17940	.62000
13	.269	.501	-.389	.39293	.21292	-.30500
14	.406	.025	.469	.01623	.32702	.37343
15	.070	.492	.320	.39133	.05590	.25240
16	.235	.652	.154	.50908	.17808	.11930
17	.332	.514	.189	.36049	.22681	.12905
18	.515	.279	.261	.19043	.36971	.18649
19	.199	.385	.432	.27043	.13518	.29513
20	.777	.026	.047	.00435	.62663	.03194
21	.281	-.036	.592	-.02446	.21368	.47611
22	.512	.261	.191	.20093	.40269	.14634
23	.046	.481	.456	.38698	.03320	.36284
24	.376	.606	.094	.49106	.30153	.06792
25	.362	.394	.357	.32074	.27943	.28106
26	.345	.385	.193	.30295	.27599	.14752
27	.448	.066	.165	.06036	.35028	.12648
28	-.093	.310	.428	.24168	-.07186	.33530
29	.026	.682	-.130	.51805	.02410	-.10475
30	.463	.254	.130	.18770	.33336	.10031