Construct Equivalence on Translated Achievement Tests

Mark J. Gierl

To permit valid group comparisons on translated tests, the construct measured by these tests must be equivalent across language groups. The purpose of this study was to empirically assess the construct equivalence of translated achievement tests in mathematics and social studies. These tests were administered to three groups of Grade 6 examinees in Canada: English, French Immersion, and Francophone students. Results from a principal components and confirmatory factor analysis supported the assumption that the constructs are unidimensional. However, results from the multi-sample analysis indicated that the factor structure and error variances are not equivalent across content areas.

L'étude avait pour but de vérifier empiriquement l'équivalence linguistique du construit de tests de connaissances en mathématiques et en sciences humaines. Ces tests ont été administrés à trois groupes d'élèves de sixième année au Canada : anglophones, en immersion française et francophones. Les résultats d'une analyse des composantes principales et des facteurs de confirmation soutiennent l'hypothèse selon laquelle les construits sont unidimensionnels. Toutefois, les résultats de cette analyse à échantillons multiples ont indiqué que la structure factorielle et les variances des écarts ne sont pas équivalentes selon la matière envisagée.

Achievement tests are often adapted or translated for use in different languages and cultures. Many examples exist. At the international level of testing, the International Association for the Evaluation of Educational Achievement (IEA) conducted the Third International Mathematics and Science Study (TIMSS) in 1995 by administering tests in 42 different languages in 60 participating countries. These are considered to be high-stake achievement tests because evaluators use results to compare student performance and to evaluate the effectiveness of educational policies and practices across the participating countries. At the local level of testing, Alberta Learning, to cite one example, has translated 8 of their 11 English high-school exiting exams into French to fulfill obligations under section 23 of the *Canadian Charter of Rights and Freedoms,* which recognizes the right of Francophones to educate their children in French. These achievement tests are also considered high-stake because the test scores contribute 50% towards students' final course grades, which, in turn, are used for postsecondary entrance and scholarship decisions. These trends are expected to continue. Hambleton (1993, 1994; also see Sireci, 1997) speculates that test adaptations and translations will become more prevalent in the future because of increased international testing, increased demand for credentialling and licensure exams in multiple languages, and growing interest in cross-cultural research.

Test translation is an important measurement topic since the validity of scores on any translated achievement test depends, in part, on the accuracy of the test adaptation. This topic is especially important in a bilingual country like Canada because many tests are administered in both official languages (English and French). A poor translation can change the validity for one set of test scores and adversely influence their comparability, meaning, and interpretability if the construct measured by the two forms is not equivalent – the construct being a theoretical representation of the underlying trait, concept, attribute, processes, or structures the test is designed to measure (Messick, 1989). In any study designed to compare examinees from two or more language groups or cultures, the construct measured by the tests must be equivalent if the comparison is to be meaningful (Gierl, Rogers, & Klinger, 1999; Hambleton, 1994; Hulin, 1987; van de Vijver & Hambleton, 1996; van de Vijver & Poortinga, 1997). If the same construct is measured in two or more language groups or cultures, then the tests are construct equivalent. Alternatively, if a different construct is measured in two or more language groups or cultures, then the tests are not construct equivalent (van de Vijver & Leung, 1997). Construct equivalence must hold in any testing situation in which test developers or users wish to compare and properly interpret the scores of different groups of examinees; it is a fundamental assumption.

Construct equivalence also has implications for the proper use of psychometric procedures used to develop and evaluate translated and adapted tests. For example, item response theory (IRT) is a popular psychometric approach for developing multilingual tests (van de Vijver & Leung, 1997), evaluating translation differential item functioning (e.g., Budgell, Raju, & Quartetti, 1995; Ellis, 1989; Hambleton, 1994; Hambleton & Kanjee, 1995; van de Vijver & Leung, 1997), and equating scores across language and cultural groups (Angoff & Cook, 1988). IRT provides a mathematical approach to modelling the relationship between an unobservable latent trait, theta, and the probability that an examinee with a given theta will answer an item correctly (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). Within the IRT framework, developers have created numerous models and applied them to practical testing problems. However, many current IRT applications using achievement test data focus on unidimensional models in which the relationship between the latent trait and item performance is modelled using the one-, two-, or three-parameter logistic function. In these applications, it is assumed that a *single latent trait* underlies test performance *and* that it is the same unidimensional construct across groups, thereby permitting meaningful comparisons. Once again, construct equivalence is a crucial assumption.

The purpose of the research I present in this article was to empirically assess the construct equivalence of two translated achievement tests. These tests were administered to three different groups of examinees – English, French Immersion, and Francophone students – in the content areas of mathematics and social studies. The data were analyzed using principal components analysis and confirmatory factor analysis. This type of study is particularly useful because researchers have for the most part overlooked the issue of construct equivalence. Van de Vijver and Leung (1997) noted that despite the importance of establishing construct equivalence when comparing different language groups and conducting sophisticated but increasingly popular psychometric analyses using unidimensional item response theory, "examples of confirmatory factor analyses in cross-cultural studies are not numerous" (p. 102). My study also provides an example of how researchers and practitioners can evaluate construct equivalence in the achievement domain.

METHOD

Student Samples and Achievement Tests

Data were collected from 1,286 Grade 6 students (500 English; 500 French Immersion; 286 Francophone) who wrote the 1997 administration of a mathematics achievement test and 1,284 Grade 6 students (500 English; 500 French Immersion; 284 Francophone) who wrote the 1997 administration of a social studies achievement test. Data for the English and French Immersion students were randomly selected from a database of approximately 38,000 English and 2,800 French Immersion students. Data for all Francophone students were used. The achievement tests were administered in the province of Alberta. All students in the province are expected to write the achievement tests and participation rates exceed 95% of the student population for any given year. Test scores do not contribute to students' final course grades although teachers are encouraged to mark the tests and use the results for student grading. Alberta Learning administers these achievement tests in both English and French.

The three groups considered in this analysis – English, French Immersion, and Francophone students – are unique but comparable. Because Canada has two official languages, different language and cultural groups can be identified in many school districts. In this study, English-speaking examinees represent the dominant language and cultural group since the majority of students receive instruction in this language at English-speaking schools. English-speaking students are tested in English.

French Immersion students are in programs where the language of instruction is French. Immersion programs are typically embedded in English-speaking schools. The Immersion program is designed for students whose first language is not French but who want to become functionally fluent in French and to develop an understanding and appreciation of French culture in addition to mastering English. Thus, French Immersion students are linguistically distinct from English-speaking students. Immersion students are tested in French.

Francophone students are also in programs where French is the language of instruction but these students attend French schools. In Canada, French schools exist for students who have at least one French-speaking parent because it is believed (and supported in Section 23 of the *Canadian Charter of Rights and Freedoms*) that students whose first language is French have linguistic, educational, cultural, and personal identity needs different from those of students learning French as a second language, such as Immersion students. To meet these needs, French schools exist where one objective of instruction is full mastery of French as a mother tongue and the establishment of a sense of identity and belonging to the French community. Hence, Francophone students are to some extent culturally distinct from French Immersion students and both these groups are linguistically and culturally distinct from English-speaking students. Alberta Learning tests Francophones in French, using the same examinations given to the French Immersion students.

In this study, construct equivalence was evaluated in two content areas: mathematics and social studies. The mathematics achievement test contained 50 multiple-choice items, with each item having four options. Items in the test specifications for mathematics are classified by Alberta Learning test developers into five curricular content areas (number relations, fractions, computation and operations, measurement and geometry, and data analysis) and two cognitive areas (knowledge and skills). The social studies achievement test contained 49 multiple-choice items (the original test contained 50 items but one item was dropped because of a translation error); each item had four options. Items in the test specifications for social studies are classified by test developers into three curricular content areas (local government, ancient Greece, and China) and two cognitive areas (knowledge and skills). Questions in both tests are based on concepts, topics, and facts in the provinicial Program of Studies (Alberta Learning, 1989, 1996).

A committee of item writers and a test development specialist developed all items in English. Alberta Learning then translated the items into French using a four-step process. First, the items were translated from English into French by one translator during item development. This translator referenced the Program of Studies and approved textbooks for grade-level and subject-specific terminology. Second, a committee comprising at least one French Immersion and one Francophone teacher along with a bilingual test developer validated the translated test. In this step, the validation committee determined the comparability of the English and French versions of the test by comparing the two tests, referring to the Program of Studies and to appropriate textbooks. Once the committee had reviewed the test, the translator and test developer received comments and feedback on the accuracy and appropriateness of the translated test. Third, the test developer, acting on the committee's recommendations, decided on final changes. The translator made these changes. Fourth, both the test developer and the achievement testing unit director reviewed and finalized the translated test. The translator in this process was a teacher with 23 years of experience in English-French translation.

Statistical Analyses

To assess the construct equivalence of the tests, parcels were used. A parcel is the sum of two or more items that serves as the unit of analysis in a confirmatory factor analysis. Parcelling has many advantages. For example, it tends to produce indicators distributed normally – a key assumption for maximum likelihood parameter estimation. It also results in stronger indicators with increased reliability and decreased error variance. Finally, parcel analyses can be used with smaller sample sizes than item-level analyses. Parcelling also has disadvantages: information about individual items is lost, parcels must be relatively unidimensional, and the resulting parameter estimates depend on the items assigned to each parcel.

To capitalize on the advantages of this procedure, parcels were created by summing items in each curricular content area by cognitive levels cells in the test specifications. The specifications guided test construction and provided the test developer's representation of the content areas and cognitive skills measured by the exam. Test specifications are readily available for researchers and practitioners; they are easy to use; and they can guide the substantive interpretation of the analysis. Moreover, test developers judge items assigned to each cell in the test specifications to be similar in content and cognitive coverage and, hence, relatively homogenous. In the current study, mathematics had five curricular content areas and two cognitive levels, resulting in 10 parcels. Social studies had three content areas and two cognitive levels, resulting in six parcels. These parcels served as the unit of analysis in the evaluation of construct equivalence using two methods.

First, I conducted a principal components analysis using the Pearson product-moment correlation matrices for the English, French Immersion, and Francophone examinees on the mathematics and social studies parcels. If a test is unidimensional, the eigenvalue for factor one will be appreciably larger than the eigenvalues for the remaining factors. This analysis was included because researchers often use this approach to evaluate the dimensionality of achievement test data (Hambleton, Swaminathan, & Rogers, 1991).

Second, I conducted a confirmatory factor analysis. This kind of analysis provides a more rigorous assessment of the latent structure across groups than does a principal components analysis because the researcher must specify an identified initial model, and this model is tested directly. A one-factor model was fit to the English, French Immersion, and Francophone data for mathematics and social studies achievement tests using LISREL 8.14 with maximum likelihood estimation (Jöreskog & Sörbom, 1996). Only the one-factor model was assessed because the tests are assumed to be unidimensional, and are interpreted in this manner for all test uses and applications (i.e., students, teachers, and parents receive only students' total test scores).

In addition, I conducted a multiple-sample analysis to evaluate the equivalence of the factor structures, factor loadings, and error variances across the English, French Immersion, and Francophone samples using the mathematics and social studies achievement tests (Byrne, Shavelson, & Muthen, 1989; Jöreskog, 1971; Jöreskog & Sörbom, 1996). Three nested models were sequentially tested by equating the number of factors, factor loadings, and error variances across the three groups. Pearson product-moment correlation matrices were used in the confirmatory factor analyses.

Confirmatory factor analytic models are assessed, in part, using goodness-of-fit indices. Currently, there is little agreement on which index provides the best answer to the question of model fit (e.g., Bollen & Long, 1993; McDonald & Marsh, 1990; Mulaik, James, Van Alstine, Bennett, Lind, & Stilwell, 1989). Consequently, I used three types of fit indices to assess each model. The first index is the chi-square statistic, which determines if the restrictive hypothesis tested can be rejected. A model is considered to have acceptable fit if the difference between the variance-covariance matrix generated by the original data and by the hypothesized solution is small, yielding a nonsignificant chi-square. The chi-square statistic is dependent on sample size and often results in a statistically significant difference when large samples, like those in the current study, are used, even when fit appears good using other indices. Despite this limitation, I included the chi-square because it is one of the most frequently used fit indices in a structural analysis for educational research (Gierl & Mulvenon, 1995). The chi-square statistic is also useful, at times, because it is the basis for other measures of model fit.

The second index I used is the root mean square error of approximation (RMSEA). The RMSEA provides a measure of fit that adjusts for parsimony by assessing the discrepancy per degree of freedom in the model. That is, RMSEA takes into account the number of free parameters required to achieve a given level of fit. Based on their practical experience, Browne and Cudek (1993) suggest that "a value of the RMSEA of about 0.05 or less would indicate a close fit of the model in relation to the degrees of freedom" (p. 144).

The third index is the standardized root-mean square residual (RMR), which represents an average of the absolute discrepancies between the observed correlation matrix and the hypothesized correlation matrix, and serves as a type of goodness-of-fit measure. A small RMR, generally 0.05 or less, indicates good fit.

RESULTS

Psychometric Characteristics of the Test Forms, Items, and Parcels

The observed psychometric characteristics of the mathematics and social studies tests for the English- and French-speaking examinees is summarized in Table 1. Typically, the kind of differences reported in Table 1 are tested between groups. However, the large samples in this study result in many differences that are statistically but not practically significant. Hence, I will not report statistical outcomes but rather highlight some general trends. First, the psychometric characteristics of the items are comparable for the English- and French-speaking examinees. The measures of internal consistency, difficulty, and discrimination are quite similar in mathematics and social studies for all three groups. Second, in mathematics, the French Immersion examinees received the highest mean score; in social studies, the English-speaking examinees received the highest mean score. However, for both tests, across all three groups the effect sizes associated with these mean differences are relatively small. Third, the standard deviations, skewness, and kurtosis for each test are similar for the three groups.

I also computed summary statistics for the mathematics and social studies parcels. These are presented in Table 2. The mean correlation across item parcels in mathematics ranged from .33 to .38, indicating that, on average, the parcels were positively correlated with one another. The mean

TABLE 1

Psychometric Characteristics for the English and French Versions in Mathematics and Social Studies

	Mathematics			*Social Studies*		
Characteristic	*EN*	*FI*	*FR*	*EN*	*FI*	*FR*
No. of examinees	500	500	286	500	500	284
No. of items	50	50	50	49	49	49
No. of words	2713	3066	3066	3354	4157	4157
Mean	35.99	37.16	35.24	33.4	31.92	29.93
SD	8.06	7.4	8.01	8.31	7.84	7.84
Skewness	–.60	–.60	–.35	–.48	–.32	–.10
Kurtosis	–.16	–.24	–.61	–.31	–.68	–.66
Internal consistency[a]	.88	.86	.87	.87	.85	.84
Mean item difficulty	.72	.74	.70	.68	.65	.61
SD item difficulty	.15	.15	.15	.11	.12	.13
Range item difficulty	.24–.91	.24–.95	.20–.91	.40–.88	.39–.86	.35–.85
Mean item discrimination[b]	.47	.44	.44	.44	.39	.37
SD item discrimination[b]	.14	.13	.14	.12	.14	.14
Range item discrimination[b]	.05–.79	.09–.74	.06–.72	.10–.67	.11–.63	.03–.67

Note. EN is English, FI is French Immersion, and FR is Francophone.
[a] Cronbach's alpha. [b] Biserial correlation.

correlation in social studies was even higher, ranging from .44 to .53, indicating that the item parcels were relatively homogeneous.

Latent Structure From Principal Components Analysis

Results from the principal components analysis indicate that test data contain a dominant first factor across all three groups for both the mathematics and social studies achievement tests. For the English form of the mathematics test, the eigenvalues for factors 1 through 5 were 3.97, 0.28, 0.25, 0.16, and 0.14. The eigenvalue for the first factor on the English form was 14.2 times larger than the eigenvalue for the second factor, whereas the eigenvalues for the second and third factors were not distinguishable. For the French form of the mathematics test with the French Immersion sample, the eigenvalues for factors 1 through 5 were 3.55, 0.34, 0.20, 0.15, and 0.13. The eigenvalue for the first factor was 10.4 times larger than the eigenvalue for the second factor, whereas the eigenvalues for the second and third factors were, again, hardly distinguishable. For the French form of the mathematics test with the Francophone sample, the eigenvalues for factors 1 through 5 were 3.94, 0.33, 0.26, 0.21, and 0.19. The eigenvalue for the first factor was 11.9 times larger than the eigenvalue for the second factor, and the eigenvalues for the second and third factors were not distinguishable.

Results for the social studies achievement tests are similar. For the English form, the eigenvalues for factors 1 through 5 were 3.63, 0.62, 0.52, 0.45, and 0.40. The eigenvalue for the first factor was 5.9 times larger than the eigenvalue for the second factor, whereas the eigenvalue for the second and third factors were more similar. For the French form of the social studies test with the French Immersion sample, the eigenvalues for factors 1 through 5 were 3.43, 0.65, 0.53, 0.51, and 0.48. The eigenvalue for the first factor was 5.3 times larger than the eigenvalue for the second factor, whereas the eigenvalues for the second and third factors were not distinguishable. For the French form of the test with the Francophone sample, the eigenvalues for factors 1 through 5 were 3.23, 0.69, 0.58, 0.53, and 0.52. The eigenvalue for the first factor was 4.7 times larger than the eigenvalue for the second factor; the eigenvalues for the second and third factors were, again, quite similar.

The results from the principal components analysis provide evidence to support the unidimensional assumption for the mathematics and social studies data across the English and French forms of the test. The eigenvalues for the first factor in both content areas and across all three groups were appreciably larger that the eigenvalues for the second factor.

TABLE 2

Summary Statistics for the Item Parcels

	Mathematics			*Social Studies*		
Summary statistic	*EN*	*FI*	*FR*	*EN*	*FI*	*FR*
Mean correlation across parcel	.38	.33	.36	.53	.49	.44
Standard deviation across parcel	.08	.09	.11	.05	.05	.05
Correlation range within parcel	.23–.59	.12–.56	.14–.64	.46–.62	.40–.57	.34–.53

Note. EN is English, FI is French Immersion, and FR is Francophone.

TABLE 3

Fit Indices for the One-Factor Model Across Content Areas as a Function of Language Group

	χ^2			*df*			*RMSEA*			*RMR*		
Content area	*EN*	*FI*	*FR*	*EN*	*FI*	*FR*	*EN*	*FI*	*FR*	*EN*	*FI*	*FR*
Mathematics	71.63*	88.88*	57.14	35	35	35	.046	.056	.047	.029	.036	.034
Social studies	25.40	15.85	8.99	9	9	9	.060	.039	.000[a]	.024	.020	.020

Note. EN is English, FI is French Immersion, and FR is Francophone.
[a]Because the degrees of freedom exceed the chi-square value, RMSEA has been set to zero (see Browne & Cudek, 1993).
*p < .01.

Latent Structure From Confirmatory Factor Analysis

I used confirmatory factor analysis to evaluate the factor structure and parameter equivalence in mathematics and social studies across the English, French Immersion, and Francophone samples. Results from the confirmatory factor analysis in both content areas support the unidimensional assumption for the data. The one-factor model provided good fit to the English, French Immersion, and Francophone data on both the mathematics and social studies achievement tests, as shown in Table 3. Although chi-square values were statistically significant for the English and French Immersion samples on the mathematics achievement tests, the RMSEA and RMR, two indices of model fit less sensitive to sample size than the chi-square statistic, were small across all three groups, indicating good model fit. For social studies, the chi-square value was statistically significant for the English sample but not for the French Immersion and Francophone samples. Moreover, the RMSEA and RMR were small across all three groups, indicating good model fit.

Although the one-factor model adequately fit all three groups in both content areas, the results from the multiple-sample analysis suggest that parameters in this model are not equivalent across the three groups in mathematics. The results of the multiple-sample analysis are presented in Table 4. The one-factor model was fit separately for the English, French Immersion, and Francophone samples, and a chi-square statistic was computed to assess parameter invariance for each model across the three groups. Three nested models were sequentially tested by equating the number of factors, factor loadings, and error variances across the three groups. For the mathematics test, models 1 and 2 and models 2 and 3 were statistically different, indicating that the three groups did not have comparable factor loadings and error variances. Although the RMSEA and RMR for model 1 were small, indicating strong model fit, the RMSEA and RMR for models 2 and 3 were larger, indicating poor model fit across the groups when the factors loadings and error variance were equated. These outcomes suggests that a unidimensional construct adequately describes the student-response data in mathematics across the three groups but the structure of this construct (i.e., the factor loadings and error variances) is not equivalent across the groups.

The results from the multiple-sample analysis in social studies suggest that parameters in the one-factor model are equivalent across the three groups. Models 1 and 2 and models 2 and 3 were not statistically different, indicating that the factor loadings and error variances were similar across the three groups. Moreover, the RMSEA and RMR were small for models 1, 2, and 3, indicating good fit.

TABLE 4

Tests for Model Equivalence Between English, French Immersion, and Francophone Examinees Across Content Areas

Content Area	χ^2	*df*	*RMSEA*	*RMR*
Mathematics				
Model 1	217.64*	103	.051	.034
Equated number of factors				
Model 2	298.56*	123	.058	.077
Equated number of factors				
Equated factor loadings				
Model 3	382.61*	143	.063	.065
Equated number of factors				
Equated factor loadings				
Equated error variances				
Social Studies				
Model 1	47.02*	25	.045	.016
Equated number of factors				
Model 2	50.31	37	.029	.027
Equated number of factors				
Equated factor loadings				
Model 3	67.20	49	.029	.034
Equated number of factors				
Equated factor loadings				
Equated error variances				

*$p < .01$.

TABLE 5

Model Comparison for Mathematics and Social Studies

Model Comparison	χ^2	*df*
Mathematics		
Model 1 vs. Model 2	80.92*	20
Model 2 vs. Model 3	84.05*	20
Social Studies		
Model 1 vs. Model 2	3.29	12
Model 2 vs. Model 3	16.89	12

*$p < .01$.

In short, the confirmatory analyses in mathematics and social studies suggest the one-factor model adequately describes the parcel data across the three groups of examinees. However, the multi-sample analysis in mathematics indicates that the parameters for this model are not equivalent across all three groups. In contrast, the parameters in social studies do appear to be invariable across groups, using a one-factor model.[1]

DISCUSSION AND CONCLUSION

Test translation is an important measurement topic because achievement tests are increasingly being translated into different languages to permit group comparisons. In these types of studies, the construct measured by the test must be equivalent across groups to allow for meaningful comparisons. In the present study, the construct equivalence of two translated achievement tests was assessed for English, French Immersion, and Francophone examinees.

Results from the principal components analysis support the unidimensional assumption. The eigenvalues for the first factor in both content areas and across all three groups were appreciably larger than the eigenvalues for the second factor. Results from the confirmatory factor analysis support the unidimensional assumption but present a more complex picture. The one-factor model provided adequate fit to the English, French Immersion, and Francophone data on both the mathematics and social studies achievement tests. However, the results from the multiple-sample analysis suggested that parameters in the one-factor model were equivalent across groups in social studies but not in mathematics. In mathematics, model testing revealed that the three groups had comparable factors but not comparable factor loadings or error variances. As a result, group comparisons in mathematics may not be appropriate until test developers evaluate the nature of this difference. By contrast, model testing in social studies revealed that the three groups had comparable factors, factor loadings, and error variance; consequently, group comparisons are justified.

Two issues in this study warrant further investigation. First, this study provides a method for comparing constructs across groups. However, this method reveals little about the substantive meaning of these constructs. Complex differences between linguistic and cultural groups mean that members in each group may have interpreted and understood the same construct in different ways – this appears to be the case in mathematics, where the factor loadings and error variances were not equivalent across the English, French Immersion, and Francophone examinees. As well, there is growing acceptance in the measurement community that the psychology of test performance must also be understood in order to develop, score, and validly interpret results from achievement tests (e.g., Frederiksen, Mislevy, & Bejar, 1993; Gierl, 1997; Mislevy, 1996; Nichols, Chipman, & Brennan, 1995; Nichols & Sugrue, 1999; Snow & Lohman, 1989; van de Vijver, 1994). Currently, test developers know little about the cognitive processes that examinees actually use as they respond to test items on different language forms of achievement tests. To better understand the psychological meaning of achievement constructs such as mathematics and social studies, researchers and practitioners need to focus on the relations between cognition and task performance by studying students' cognitive processes as they respond to test items in different content areas in addition to assessing the structural features of these constructs. Snow and Lohman (1989) provide this reminder in their seminal chapter, "Implications of Cognitive Psychology for Educational Measurement":

> As a substantive focus for cognitive psychology then, "ability," the latent trait θ in EPM [educational and psychometric measurement] models, is not considered univocal, except as a convenient summary of amount correct regardless of how obtained. Rather, a score reflects a complex combination of processing skills, strategies, and knowledge components, both procedural and declarative and both controlled and automatic, some of which are variant and some invariant across persons, or tasks, or stages of practice, in any given sample of persons or tasks. In other samples of persons or situations, different combinations and different variants and invariants might come into play. Cognitive psychology's contribution is to analyze these complexes. (pp. 267–268)

Because many achievement constructs are inadequately described and poorly understood, measurement specialists could also pursue Snow and Lohman's challenge by studying the psychological characteristics that underlie latent variables even when they are demonstrated to be structurally equivalent across groups. My study revealed that a one-factor model describes the responses of English, French Immersion, and Francophone examinees in mathematics and social studies and that the structural components of this model are invariant across the three groups in social studies but not in mathematics. Future research is needed for us to understand the nature of these similarities and differences.

Second, the unidimensional assumption could also be assessed using higher-order factor analysis, although applications of this type are rare (Bollen, 1989; Jöreskog & Sörbom, 1996) and, when applied, seldom provide adequate fit to the data (e.g., Bagozzi & Edwards, 1998; Bagozzi & Heatherton, 1994). In this type of model, latent variables directly influencing the observed variables may also be influenced by other latent variables that do not have a direct effect on the observed variables. In the current study, a higher-order factor analytic model could be assessed using the test items as indicators for the parcels which, in turn, serve as indicators for one general factor. This model remains an alternative that future researchers should consider, as it may provide a more refined view of the underlying structure of complex constructs such as mathematics and social studies.

ACKNOWLEDGMENTS

The Social Sciences and Humanities Research Council of Canada (SSHRCC) and the Social Science Research (SSR) Fund at University of Alberta supported this research. I thank Les Hayduk for his comments on an earlier version of this manuscript. I also thank Deborah Bandalos for her insightful comments on the use of item parcels in structural equation modelling.

NOTE

1. The standardized factor loadings and the error variances in the one-factor model for the English, French Immersion, and Francophone students in mathematics and social studies are available from the author upon request.

REFERENCES

Alberta Learning. (1989). *Program of Studies: Social Studies.* Edmonton: Alberta Learning, Curriculum Standards Branch.

Alberta Learning. (1996). *Program of Studies: Mathematics.* Edmonton: Alberta Learning, Curriculum Standards Branch.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (College Board Report No. 88-2). New York: College Entrance Examination Board.

Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods, 1,* 45–87.

Bagozzi, R. P., & Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. *Structural Equation Modeling, 1,* 35–67.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bollen, K. A., & Long, J. S. (1993). *Testing structural equation models.* Newbury Park, CA: Sage Publications.

Browne, M. W., & Cudek, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage Publications.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19,* 309–321.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

Ellis, B. B. (1989). Differential item functioning: Implications for test translation. *Journal of Applied Psychology, 74,* 912–921.

Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (1993). *Test theory for a new generation of tests.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gierl, M. J. (1997). Comparing the cognitive representations of test developers and students on a mathematics achievement test using Bloom's taxonomy. *Journal of Educational Research, 91,* 26–32.

Gierl, M. J., & Mulvenon, S. (1995, April). *Evaluating the application of fit indices to structural equation models in educational research: A review of the literature from 1990 through 1994.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). *Consistency between statistical procedures and content reviews for identifying translation DIF.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, QC.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-cultural studies. *European Journal of Psychological Assessment, 9,* 57–68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229–244.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11,* 147–157.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Hulin, C. L. (1987). A psychometric theory of evaluations of items and scale translations. *Journal of Cross-Cultural Psychology, 18,* 115–142.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago, IL: Scientific Software.

Lord. F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107,* 247–255.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). Washington, DC: American Council on Education; New York: Macmillan.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379–416.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105,* 430–445.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum.

Nichols, P., & Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. *Educational measurement: Issues and practice, 18,* 18–29.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16,* 12–19.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.), (pp. 263–331). Washington, DC: American Council on Education; New York: Macmillan.

van de Vijver, F. J. R. (1994). Item bias: Where psychology and methodology meet. In A. Bouvy, F. J. R. van de Vijver, P. Boski, & P. Schmitz (Eds.), *Journeys into cross-cultural psychology* (pp. 111–126). Lisse, The Netherlands: Swets & Zeitlinger.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1,* 89–99.

van de Vijver, F., & Leung, K. (1997). *Methods and data-analysis for cross-cultural research.* Thousand Oaks, CA: Sage Publications.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13,* 21–29.