DOCUMENT RESUME

ED 382 031 FL 022 946

AUTHOR Ginther, April; Stevens, Joseph

TITLE Language Background, Ethnicity, and the Internal Construct Validity of the Advanced Placement Spanish Language Examination.

PUB DATE May 95

NOTE 27p.

PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Bilingualism; Black Students; Comparative Analysis; *Construct Validity; *Ethnicity; *Language Proficiency; *Language Tests; Second Language Learning; *Spanish; *Student Placement; Test Construction; Testing; White Students

ABSTRACT

The internal construct validity of the 1989 Advanced Placement Spanish Language Examination was analyzed to determine whether the traditional four-factor examination structure hypothesized by test developers (listening, reading, writing, speaking) was invariant for relevant subpopulations. Using a nested hierarchical design, multiple confirmatory analyses were conducted for examinees differing in language background and ethnicity. An analysis of exam structure for Latin Spanish-speaking examinees served as the starting reference to which the examination structures of Mexican Spanish-speaking, Mexican Spanish/English bilingual, White English-speaking, and Black English-speaking examinees were compared. While the four-factor examination structure fit all groups, factor loadings, variances, covariances, and variable uniqueness all differed significantly across groups. The differences found have implications for test development as well as for the use and interpretation of scores in language proficiency testing. Contains 25 references and 4 tables. (Author)

Language Background, Ethnicity, and the Internal Construct Validity of the Advanced Placement Spanish Language Examination

April Ginther

Language Background, Ethnicity, and the Internal Construct Va.idity of the Advanced Placement Spanish Language Examination

April Ginther
Educational Testing Service

Joseph Stevens
University of New Mexico

The internal construct validity of the 1989 Advanced Placement Spanish Language Examination was analyzed to determine whether the traditional four-factor examination structure hypothesized by test developers (Listening, Reading, Writing, and Speaking) was invariant for relevant subpopulations. Using a nested hierarchical design, multiple confirmatory analyses were conducted for examinees differing in language background and ethnicity. An analysis of exam structure for Latin Spanish-speaking examinees served as the starting reference to which the examination structures of Mexican Spanish-speaking, Mexican Spanish/English bilingual, White English-speaking and Black English-speaking examinees were compared. While the four-factor examination structure fit all groups, factor loadings, variances, covariances, and variable uniquenesses all differed significantly across groups. The differences found have implications for test development as well as for the use and interpretation of scores in language proficiency testing.

Background

Confirmatory Factor Analysis (CFA) was used to examine the structure of the Advanced Placement Spanish Language Examination for a Latin Spanish-speaking reference group in comparison to (1) a Mexican Spanish-speaking group, (2) a Mexican Spanish/English bilingual group, (3) a White group of Spanish foreign language learners, and (4) a Black group of Spanish foreign language learners. In the analysis of examination structure with respect to the performance of first- and foreign-language subpopulations, this study differs in several ways from most other analyses of both subpopulation performance and language proficiency.

Examinations of subpopulation performance in the testing literature are often based on differential item functioning or analyses of mean score differences. In the case of the former, only individual items are examined, and in the case of the latter, a composite is created. When only individual items or means are examined, the ability to detect patterns across groups of items with respect to relevant subpopulations is lost. Focusing on possible differences in examination structure through analysis of variances and covariances allows the researcher to address differences in the construct being measured.

In direct reference to standardized testing, Duran (1988) suggests the possibility of differences in examination structures for different groups: "Contemporary cross-cultural research suggests that there are intimate connections among the ways people perceive the nature of problem-solving situations, problem-solving tasks, and sociocultural experiences" (p. 574). The presence of such differences should be taken into account not only to ensure accurate and fair interpretation of test scores but also to develop better conceptualizations of the underlying competencies involved.

Although first-language proficiency often serves as an idealized criterion in discussions of second- and foreign-language proficiency, explicit empirical comparisons of the performance of first-, second-, and foreign-language learners are infrequent. Studies of the fit of models with particular numbers of factors have been conducted in studies of the second-language acquisition of English, but these studies do not compare first- and second-language speaker performance.

Cziko (1982) argues that meaningful interpretation of patterns of results on language tests should include analysis of relevant background variables. He argues, "If this is done, then we may well find that what is taken as evidence for either a one-factor or multi-factor working model of communicative competence may instead be simply an indication that the pattern of language proficiency one acquires is related to the type and amount of exposure to the language that one has" (p. 7). A step toward clarifying the effects of amount and type of exposure on language proficiency is to determine whether systematic differences in examination structure are related to background variables.

Review of Related Factor-Analytic Studies

Factor-analytic studies concerning the performance of examinees on language proficiency tests involve the demonstration of how components are psychologically interrelated to form factors. Much of this research has centered on the question of whether and to what extent examination structure is divisible into separate factors. Appropriate division of components of an examination has implications for scoring, placement of examinees, and our understanding of the development of language proficiency.

One source of information about language proficiency examination structure is provided by research conducted on the *Test of English as a Foreign Language* (TOEFL). Swinton and

Powers (1980) conducted a factor-analytic study in which the structure of the examination was the object of analysis for seven groups of examinees differing in first-language background. They conclude that a three-factor model (listening comprehension, reading and vocabulary comprehension, and structure and written expression) is appropriate for most groups, but found the greatest distinctions in factor structure for the group with the highest proficiency (Germanic) and the least amount of factor differentiation for the language group with the lowest proficiency (Farsi).

Hale, Stansfield, Rock, Hicks, Butler, and Oller (1988) conducted confirmatory factor analyses for each of nine major language groups on the TOEFL and found that only two factors were necessary to account for performance. A model comprised of listening and non-listening factors rather than the tnree factor model provided the best fit for all groups.

In an attempt to account for the disagreement, Hale, Rock, and Jirele (1989) conducted a follow-up study which analyzed the data not only with respect to native-language background but also took proficiency and domestic versus overseas location of testing into account. The domestic versus overseas location was found to have little effect, and these populations were combined in subsequent analyses.

Hale, Rock, and Jirele found, in agreement with the the results of the Hale et al. (1988) study, that the pattern of results supported the idea of a two-factor rather than a three-factor solution. While results demonstrated that going from a one-factor to a three-factor solution improved the goodness of fit in each case, the gain in goodness-of-fit with respect to a three-factor solution was not substantial: psychometrically, with respect to signficance and the preference for the most parsimonious solution, the two-factor solution provided the best fit to the data. However, the researchers ultimately argue for a three-factor solution becuase taking proficiency into account when interpreting the results restricts the viability of a two-factor solution.

While correlational evidence supported the idea of a two-factor solution, apparently different aspects of proficiency were tapped by the three subsections of the TOEFL. Although students may have the same rank ordering in content areas, it is possible they could score below a threshold in one area but above the threshold in another area. Thus correlations between subsections would be high, but having the subsection scores provides important information. Hale et al. suggest that establishing a scale of proficiency levels within each content area would allow more accurate diagnosis and explanation of the proficiencies involved.

In an analysis of language proficiency as measured by the *Test of English as a Foreign Language* (TOEFL) as well as the *Illinois English Placement Battery* (IEPT), Fouly, Bachmar, and Cziko (1990) move away from traditional conceptualizations of language proficiency (reading, writing, listening, speaking) and propose two examination structure models which differ with respect to the presence or absence of a higher-order factor. The correlated-trait (CT) model proposes that separate traits (or factors) underlie performance and that these traits are correlated with each other. The higher-order (HO) model proposes the same factors but also posits a single higher-order factor that influences the separate traits. The three traits in both models are 1) oral-aural which involves the ability to speak and understand, 2) structure and reading comprehension which involves the ability to recognize and understand written English structures, and 3) discourse competence which involves the use of language rules to interpret the cohesion and organization of a group of utterances. The fit of both models was found equally 'good,' but the researchers did not take language background or proficiency into account.

Morgan and Mazzeo (1988) studied the 1987 Advanced Placement (AP) French Language Exam. The study was undertaken to compare the relations among the listening, reading, writing, and speaking components of the exam across four populations using a series of confirmatory factor analysis models. The first two populations were AP French Language examinees who had no out-of-class exposure to French, the third consisted of examinees who had spent time in a French-speaking country, and the fourth consisted of third-year French students with no out-of-class French exposure who were enrolled in university French courses. Six models were tested ranging from a one-factor model to a six-factor model in which short and long listening items comprised separate factors along with structure, reading, writing, and speaking factors.

The AP French Language Examination differs from the TOEFL in several ways. Along with multiple choice items based on taped dialogues, narratives, and a lecture, and the multiple-choice items based on reading comprehension, vocabulary, and structure, the exam also has two constructed response sections. The first section includes two writing measures--an essay and a modified cloze test, and the second section measures speaking ability through evaluation of a taped story-telling exercise and a series of short answers to questions. In studies of the TOEFL, the multiple-choice structure items are associated with the essay, but the test developers of AP French have posited a structure in which the structure items load on the reading factor.

In contrast to studies conducted using the TOEFL, the AP French Exam was found to measure at least four dimensions associated with listening, reading, writing, and speaking proficiency. This is due, in part, to the additional speaking section of the AP French Exam in contrast to the TOEFL. Morgan and Mazzeo did not test the model that would have been roughly comparable to the two-factor TOEFL model--listening, speaking, and a general factor associated with all writing, reading, and structure items.

Interestingly, while a four-factor model was found to fit the data well, Morgan and Mazzeo found evidence for two additional factors possibly associated with item format. Goodness of fit indices were improved when two separate, but highly correlated, listening factors were included in the model--short listening and long listening. Items based on short listening passages were found more highly correlated with other parts of the exam than were the items based on longer listening passages. The items based on long listening passages were found to be more highly correlated with the longer reading comprehension passages. They speculated that the demand made to answer items based on longer listening and reading passages may uniquely tap the ability to retain information in memory. The presence of comparable factors on the current TOEFL cannot be investigated as no long-listening items exist.

Finally, group membership did not have differential impact on exam structure. However, the group populations with out-of-class French exposure did evidence slightly different scales with differing degrees of measurement precision when compared to the standard groups. It may be the case that the differences in target language exposure were not sufficient to produce differences in levels of proficiency.

These studies suggest that background characteristics of examinees do exert an influence on the internal construct validity of language exams. In the first three studies, group membership based on first-language background was examined. While results with respect to the structure of the TOEFL are inconclusive, these results caution against positing a single examination structure appropriate for all examinees. The Fouly et. al. study differed in that its purpose was to offer interpretations of the nature of language proficiency rather than to validate a particular test structure. Given findings of other studies involving the TOEFL, it appears that such attempts are problematical without consideration of learner background characteristics. One of the problems with understanding such influences is the obvious and natural confound that arises in relation to proficiency. If, however, we accept the view of communicative competence championed first by Hymes (1972) and elaborated by others (Canale and Swain, 1980; Savignon, 1983; Stern, 1983), which emphasizes the development of second- and foreign-language proficiency as a dynamic process, then attempts to understand the interactions between background variables and proficiency become central concerns in second language acquisition research.

Examination of the Advanced Placement Spanish Language Examination offered a unique opportunity to address some of these issues. The structure of AP Spanish is like that of AP French, but the population tested has very different characteristics. Unlike the TOEFL and AP French, where no first-language speakers were identified or expected to take the exam, the population of examinees for the AP Spanish exam consists of both first-language speakers and foreign-language learners of Spanish. This situation reflects the presence of Spanish-English bilingualism in a portion of the population of students who take the AP

Spanish Exam and offers the opportunity to examine whether differences in examination structure exist given the different characteristics of the subpopulations involved.

Method

The AP Spanish Language Examination is designed to assess high school students' proficiency in Spanish in order to obtain college credit or placement into advanced courses. The examination is composed of two sections: multiple choice and free response. The multiple choice section consists of 90 items, divided into four parts: listening comprehension, vocabulary, recognition of grammatical structures, and reading comprehension. The free response section is divided into two sections intended to test "the active skills of speaking and writing" (College Board, 1989, p. 3). and includes essay, modified cloze, story-telling, and directed-response tasks. The examination is conceptualized by the test developers as measuring four broad language skills: Listening, Reading, Writing, and Speaking.

Samples were drawn from a population of 9,556 examinees who took both the Scholastic Achievement Test (SAT) and AP Spanish Language Examination. Information provided by examinees on the SAT student descriptive questionnaire was used to identify examinee ethnicity and the examinee's preferred language. Random samples were drawn from each ethnic/language combination such that a maximum of 500 examinees, when available, were chosen. Samples were not analyzed if there were fewer than 200 examinees.

This sampling procedure resulted in the following groups used in the analyses: Latin-Spanish speaking (N=500), Mexican-Spanish speaking (N=307), Mexican-Bilingual (N=308), White-English speaking (N=500), and Black-English speaking (N=249). The largest Spanish speaking group, Latin-Spanish, was used as the reference group for purposes of model comparison.

Prior to analysis, multiple-choice items were combined into item parcels in order to increase reliability and to ameliorate difficulties that occur in the factor analysis of dichotomously scored items (Byrne, 1989; Dorans & Lawrence, 1987). Item parcels were created so that parcel means, standard deviations, and item types were approximately equivalent across parcels within a particular factor.

Test development documentation and program descriptive materials (College Board, 1989) were used to specify an a priori model of examination structure consisting of four factors: Listening, Reading, Writing, and Speaking. Goodness of fit of this model to the data for the Latin-Spanish group was tested with Confirmatory Factor Analysis (CFA) using LISREL 7 (Jörkeskog & Sörbom, 1988). A null model was also applied to provide a baseline for model comparisons using the Tucker-Lewis index (Marsh, Balla, & McDonald, 1988). In each comparison of groups, the following model features were held invariant from the Latin-Spanish reference group to the target comparison group with each additional invariance constraint added to the previous constraints in the hierarchy: 1) number of factors, 2) magnitude of factor loadings, 3) magnitude of factor variances, 4) magnitude of factor covariances, and 5) magnitude of variable uniquenesses. Chi-square values for the preceding model were subtracted from the chi-square value for the model of interest to provide tests of group differences for the particular parameters of interest.

Results

The four-factor model provided a good fit to the data for the Latin-Spanish group, $\chi^2(59) = 71.89$, $p > .05$, $\chi^2/\text{df} = 1.22$. A χ^2/df ratio below 2.00 is generally accepted as an indication of good model fit (Byrne, 1989) as is a nonsignificant chi square. Furthermore, the observed adjusted goodness of fit index (AGFI) of .97 approached unity, and the Tucker-Lewis Index of .99 indicated that relative to the null, the four-factor model fit the data well. Given the goodness of fit of the four-factor model for the Latin-Spanish group, a series of invariance tests were then conducted to determine whether the model fit equally well across the remaining ethnic and language groups.

Comparisons of model structure from the Latin-Spanish to the Mexican-Spanish groups showed no difference in the number of factors in the model but did show significant differences in the factor loadings ($\chi^2[9] = 23.34$, $p < .05$), factor variances ($\chi^2[4] = 23.99$, $p < .05$), and variable uniquenesses ($\chi^2[13] = 48.53$, $p < .05$). No significant differences in factor covariances were found. The same pattern of differences was found in the comparison of the Lat in-Spanish group to the Mexican-Bilingual group, except that group differences were somewhat larger.

Significant differences in model fit were found in comparisons of the Latin-Spanish group with both English-speaking groups as well. In comparing the Latin-Spanish group to the White-English group, differences were present on factor loadings ($\chi^2[9] = 35.38$, $p < .01$), factor variances ($\chi^2[4] = 544.32$, $p < .01$), factor covariances ($\chi^2[4] = 52.87$, $p < .01$), and variable uniquenesses ($\chi^2[13] = 1048.38$, $p < .01$). Comparisons of the Latin-Spanish group with the Black-English group showed very similar results with significant differences found in all tests other than the number of factors: factor loadings ($\chi^2[9] = 36.19$, $p < .01$), factor variances ($\chi^2[4] = 361.41$, $p < .01$), factor covariances ($\chi^2[4] = 56.23$, $p < .01$), and variable uniquenesses ($\chi^2[13] = 1002.72$, $p < .01$). Thus, in both sets of comparisons of the English-speaking groups with the Latin-Spanish reference group, differences were larger than those observed in comparisons to the Spanish-speaking or bilingual examinee groups, and in addition to differences in factor loadings, factor variances, and variable uniquenesses, differences were also found in factor covariances.

These results show a progression of increasing differences in examination structure from the Mexican-Spanish to the Mexican bilingual to the two English-speaking groups of examinees. These differences can be characterized by higher performance and lower variability on all of the factors by the Spanish-speaking groups. The Speaking factor, however, stands out. Significantly lower correlations of this factor with the remainder of the examination was an obvious difference in examination structure for the Spanish-speaking groups; that is, the Speaking factor was relatively unrelated to other factors for the Spanish-speaking groups, but highly related to the other factors for White and Black English-speaking groups. Factor loadings for the Spanish-speaking groups were higher on Reading, Writing, and Speaking than for the English-speaking groups. The intercorrelations of the four factors, on the other hand, were generally higher for the English-speaking groups.

Discussion

The differences in factor covariances found in this study suggest that differences in exposure to the target language may affect the development of proficiency in important ways. While factor covariances were the same in comparisons between the Latin and Mexican groups, all of whom were highly proficient speakers of Spanish, differences were found in the comparisons between the Latin and English-speaking groups. The Speaking factor was, relatively speaking, unrelated to the other factors for the Latin examinees while it is strongly related to the other factors for the White and Black groups.

Obviously, differences in out-of-class experience have a great impact in the development of language proficiency. The membership of the Spanish-speaking examinees in a language community in which Spanish is used for communicative purposes that go beyond classroom experience is reflected not only in the higher means of these groups but also in the lack of relation of the Speaking Factor to the others.

The current emphasis on communicative activities in second language theory and teaching is, in part, a reflection of the attempt to create classroom situations that more closely match students' experience in communities where the target language is spoken. Foundational to such perspectives is the belief that language proficiency is developed only through the use of the language for meaningful, communicative purposes. Correspondingly, most people who have learned or taught a second or foreign language believe that in order to develop oral proficiency or communicative competence in conversation, out-of-class experience in a setting where the target language is spoken is not merely beneficial but necessary. Apparently, when the only exposure a student has to a target language is in the classroom, the abilities developed are constrained by the academic milieu. Indeed, the students who have only an English-speaking background are also members of a well-defined language community, but the community is one which is largely characterized by what is commonly called "book learning" or academic discourse.

The generally weaker relations among factors for the Latin and Mexican Spanish-speaking examinees suggests that Listening, Reading, Writing, and Speaking are more distinct and less interdependent for native speakers. Conversely, the stronger relations among factors for non-native speakers suggests greater commonality among factors. The pattern of factor loadings for Reading, Writing, and Speaking across native and non-native Spanish-speaking groups appears to support this interpretation. When the factor loadings for Reading, Writing, and Speaking for the Latin Spanish-speaking examinees are compared to the loadings for the White and the Black English-speaking examinees, the loadings for the Latin Spanish-speaking examinees are higher in every case. This pattern suggests greater salience of these factors for native speakers as indicated by stronger relations between the measured and latent variables. Global or "general language proficiency" configuration of skills is more likely to be characteristic of language learners at lower levels of development. As either native or non-native speakers of a second or foreign language reach high levels of proficiency, it may be that the factors involved become more distinct.

In light of all the evidence, stronger relations among factors appear characteristic of lower levels of proficiency, and weaker relations among the factors along with stronger factor loadings appear characteristic of higher levels of proficiency. These results support the findings of Swinton & Powers (1980) who found greater dimensionality for the higher proficiency examinees on the TOEFL.

The results of this study demonstrate that the oral proficiency of native speakers is both quantitatively and qualitatively different from that of students who have acquired or learned what they know of the target language in class. This finding has important implications for the valid use and interpretation of these test scores. As mentioned, the AP Spanish Language Examination is used to place high school students into or out of the first two years of college level Spanish language classes. In this case, the subsections of the test are weighted when forming the total score, and the weights are calculated so that each section of the test is given equal weight in the formation of the composite. While the speaking factor is unrelated to the other factors for examinees who have out-of-class experience with Spanish, Speaking scores are given equal weight in relation to the other subsections of the tests. Therefore, some of the examinees are not being evaluated on what they have learned in class but what they have brought to class. On the other hand, the students who have not had out-of-class language experience are being evaluated in terms of their in-class experience--the only experience they have had. Using composites in this case is misleading. Because the validity of an exam is largely based on the use and interpretation of test scores, the findings of this study demonstrate that group membership should be taken into account.

Several alternatives to current practices could be employed. The scores could be weighted differently for examinees with Spanish-speaking backgrounds. Because it appears to be the case that the Speaking portion of the examination is too easy for these examinees, perhaps the composite should not include the Speaking section for examinees with a Spanish-speaking background. Furthermore, their lower scores on Listening, Reading, and Writing suggest that these students might benefit from continuing to study Spanish but not at an introductory level. The use of the examination for placement purposes rather than credit could be a solution.

An alternative would be to give examinees with different backgrounds different kinds of exams. This practice is already followed in English language testing. Native speakers of English are not given the same kinds of exams as non-native speakers. This does not mean that English should be tested in one way or another but only that we recognize differences in the test takers.

With respect to theoretical models of language proficiency or communicative competence, more large-scale studies involving both native and non-native speakers need to be conducted. Although native language proficiency is the idealized criterion in most language testing situations, these populations are seldom compared. An interesting follow-up to this study would be to examine the relations among factors for native speakers of English on a test of English as a foreign language. If the lack of relationship between Speaking and other factors were reproduced in a study with another population and another language, we might be able to add further insight into discussions of the influences of ethnicity and language background on the development of proficiency.

References

Byrne, B. M. (1989). *A Primer of LISREL. Basic Applications and Programming for Confirmatory Factor Analytic Models.* New York, NY: Springer-Verlag.

Canale, M. (1983). On some dimensions of language proficiency. In J. W. Oller Jr. (Ed.). *Issues in Language Testing Research.* (pp. 333-342). Rowley, MA: Newbury House.

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

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In *Testing the English Proficiency of Foreign Students.* Washington, DC: Center for Applied Linguistics.

Carroll, J. B. (1980). Measurement of abilities and constructs. In *Construct Validity in Psychological Measurement: Proceedings on a Colloquium on Theory and Application in Educational Measurement.* Princeton, NJ: Educational Testing Service.

College Board. (1988). *The college board technical manual for the advanced placement program.* New York: College Entrance Examination Board.

College Board. (1989). *Advanced Placement Course Description: Spanish Language and Literature.* New York: The College Board.

College Board. (1990). *Registration bulletin: SAT and achievement tests.* New York: The College Board.

Cziko, G. A. (1982, May). *Developing models of communicative competence: Conceptual, statistical, and methodological considerations.* Paper presented at the annual convention of Teachers of English to Speakers of Other Languages. Honolulu, HI. (ERIC Document ED 226 567).

Dorans, N. & Lawrence, I. (1987). *The Internal Construct Validity of the SAT.* (ETS Research Report: RR-87-35). Princeton, NJ: ETS.

Duran, R. P. (1988). Testing linguistic minorities. In R. L. Linn (Ed.). *Educational measurement* (3d ed.). (pp. 573-587). New York: American Council on Education\Macmillan.

Fouley, K. A., Bachman, L. F. & Cziko, G. A. (1990). The divisibility of language competence: A confirmatory approach. *Language Learning 40*, 1-21.

Hale, G. A., Rock, D. A., & Jirele, T. (1989). *Confirmatory Factor Analysis of the Test of English as a Foreign Language. TOEFL Research Report 32*. Princeton, NJ: Educational Testing Service.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-93). Harmondsworth: Penguin.

Jörkeskog, K. G. & Sörbom, D. (1988). *Lisrel 7 A guide to the program and applications*. Gorinchem, The Netherlands: SPSS International.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling,1*, 5-34.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*, 391-410.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Lawrence Earlbaum.

Morgan, R. & Mazzeo, J. (1988). *A Comparison of the Structural Relationships among Reading, Listening, Writing, and Speaking Components of the AP French Language Examination for AP Candidates and College Students*. RR-89-59. Princeton, NJ: Educational Testing Service.

Oltman, P. K., Stricker, L. J. & Barrows, T. (1988). *Native Language, English Proficiency, and the Structure of the Test of English as a Foreign Language. TOEFL Research Report 27.* Princeton, NJ: Educational Testing Service.

Savignon, S. J. (1972). *Communicative Competence: An experiment in Foreign Language Teaching*. Philadelphia, PA: Center for Curriculum Development.

Savignon, S. J. (1983). *Communicative Competence: Theory and Classroom Practice*. Reading, MA: Addison-Wesley.

Schachter, J. (1990). Communicative competence revisited. In B. Harley, P. Allen, J. Cummins, & M. Swain (Eds.). *The Development of Second Language Proficiency*. Cambridge: Cambridge University Press.

Stern, H. H. (1983). *Fundamental Concepts of Language Teaching.* Oxford: Oxford University Press.

Swinton, S. S., & Powers, D. E. (1980). *Factor Analysis of the Test of English as a Foreign Language for several Language Groups. TOEFL Research Report 6.* Princeton, NJ: Educational Testing Service.

Table 1

Invariance Tests of the Four-Factor Model for Latin Spanish and Mexican Spanish Examinees

	Goodness-of-fit			Change in Goodness-of-fit		
Model	χ^2	df	χ^2/df	$\Delta\chi^2$	Δdf	$\Delta(\chi2/\mathrm{df})$
1. H:Λ,f=4	138.98	118	1.18	--	--	--
2. H:Λ	162.32*	127	1.28	23.34*	9	.10
3. H:Λ,$\Phi(\sigma^2)$	186.31**	131	1.42	23.99**	4	.14
4. H:Λ,Φ	202.00**	137	1.47	15.69	6	.05
5. H:Λ,Φ,Θ	250.53**	150	1.67	48.53**	13	.20

Note: * $p < .05$
** $p < .01$

Table 2

Invariance Tests of the Four-Factor Model for Latin Spanish and White English Examinees

	Goodness-of-fit			Change in Goodness-of-fit		
Model	χ^2	df	χ^2/df	$\Delta\chi^2$	Δdf	$\Delta(\chi^2/\mathrm{df})$
1. H:Λ,f=4	142.51	118	1.21	--	--	--
2. H:Λ	177.89**	127	1.40	35.38**	9	.19
3. H:Λ,$\Phi(\sigma^2)$	722.21**	131	5.51	544.32**	4	4.11
4. H:Λ,Φ	775.08**	137	5.66	52.87**	6	.15
5. H:Λ,Φ,Θ	1823.46**	150	12.16	1048.38**	13	6.50

Note: * $p < .05$
** $p < .01$

Table 3
Parameter Estimates and Standard Errors for Responses on Parcels by Latin Spanish-speaking and by Mexican Spanish-speaking Examinees

	Factor Loadings (LAMBDA X)				Error/ Uniqueness (THETA)
Parcel	Listen	Read	Write	Speak	
L1) LS	1.00	0	0	0	1.31(.13)
MS	1.00	0	0	0	1.97(.20)
L2) LS	.82(.06)	0	0	0	1.38(.12)
MS	1.03(.12)	0	0	0	.99(.15)
L3) LS	.90(.06)	0	0	0	1.68(.14)
MS	.92(.12)	0	0	0	2.23(.21)
R1) LS	0	1.00	0	0	.62(.04)
MS	0	1.00	0	0	.87(.08)
R2) LS	0	1.03(.06)	0	0	.86(.06)
MS	0	.63(.09)	0	0	1.04(.09)
R3) LS	0	.99(.06)	0	0	.61(.05)
MS	0	1.02(.10)	0	0	.76(.07)
R4) LS	0	1.12(.06)	0	0	.56(.05)
MS	0	1.05(.11)	0	0	.82(.07)
R5) LS	0	1.05(.06)	0	0	.75(.06)
MS	0	1.01(.10)	0	0	.75(.07)
R6) LS	0	1.06(.06)	0	0	.86(.06)
MS	0	.97(.11)	0	0	.97(.09)
W1) LS	0	0	1.00	0	2.11(.19)
MS	0	0	1.00	0	1.63(.28)
W2) LS	0	0	.93(.09)	0	1.52(.15)
MS	0	0	.66(.11)	0	2.04(.20)
S1) LS	0	0	0	1.00	.08(.02)
MS	0	0	0	1.00	.15(.02)
S2) LS	0	0	0	1.56(.05)	.08(.04)
MS	0	0	0	1.81(.20)	.01(.07)

Factor Variances and Covariances (PHI)

		Listen	Read	Write	Speak
Listen	LS	2.07(.23)			
	MS	1.40(.25)			
Read	LS	.79(.09)	.87(.09)		
	MS	.61(.10)	.63(.10)		
Write	LS	1.06(.14)	.85(.10)	1.55(.23)	
	MS	.75(.15)	.78(.11)	1.67(.34)	
Speak	LS	.14(.04)	.14(.02)	.18(.03)	.19(.02)
	MS	.13(.04)	.18(.03)	.17(.05)	.21(.03)

Note: LS=Latin Spanish; MS=Mexican Spanish.

Table 4
Parameter Estimates and Standard Errors for Responses on Parcels by Latin Spanish-speaking and by White English-speaking Examinees

		Factor Loadings (LAMBDA X)				Error/ Uniqueness
Parcel		Listen	Read	Write	Speak	(THETA)
L1)	LS	1.00	0	0	0	1.31(.14)
	WE	1.00	0	0	0	1.95(.16)
L2)	LS	.82(.06)	0	0	0	1.38(.12)
	WE	1.11(.07)	0	0	0	1.74(.16)
L3)	LS	.90(.06)	0	0	0	1.68(.14)
	WE	.97(.07)	0	0	0	2.03(.16)
R1)	LS	0	1.00	0	0	.62(.04)
	WE	0	1.00	0	0	2.06(.15)
R2)	LS	0	1.03(.06)	0	0	.86(.06)
	WE	0	.92(.05)	0	0	1.67(.12)
R3)	LS	0	.99(.06)	0	0	.61(.05)
	WE	0	.98(.06)	0	0	1.76(.13)
R4)	LS	0	1.12(.06)	0	0	.56(.05)
	WE	0	1.04(.06)	0	0	1.58(.12)
R5)	LS	0	1.05(.06)	0	0	.75(.06)
	WE	0	.82(.05)	0	0	1.93(.13)
R6)	LS	0	1.06(.06)	0	0	.86(.06)
	WE	0	.81(.06)	0	0	2.25(.15)
W1)	LS	0	0	1.00	0	2.11(.19)
	WE	0	0	1.00	0	1.42(.15)
W2)	LS	0	0	.93(.09)	0	1.52(.15)
	WE	0	0	.73(.05)	0	1.85(.13)
S1)	LS	0	0	0	1.00	.08(.02)
	WE	0	0	0	1.00	.67(.09)
S2)	LS	0	0	0	1.56(.15)	.08(.04)
	WE	0	0	0	1.26(.05)	.27(.13)

Factor Variances and Covariances (PHI)

		Listen	Read	Write	Speak
Listen	LS	2.07(.23)			
	WE	2.35(.26)			
Read	LS	.79(.09)	.87(.09)		
	WE	2.10(.20)	2.66(.28)		
Write	LS	1.06(.14)	.85(.10)	1.55(.23)	
	WE	1.85(.19)	2.60(.22)	2.70(.27)	
Speak	LS	.14(.04)	.14(.02)	.18(.03)	.19(.02)
	WE	1.63(.08)	1.65(.17)	1.64(.17)	2.70(.23)

Note: LS=Latin Spanish; WE=White English.