The Use of Generalizability (G) Theory in the Testing of Linguistic Minorities

Guillermo Solano–Flores, *University of Colorado, Boulder*, and Min Li, *University of Washington, Seattle*

We contend that generalizability (G) theory allows the design of psychometric approaches to testing English-language learners (ELLs) that are consistent with current thinking in linguistics. We used G theory to estimate the amount of measurement error due to code (language or dialect). Fourth- and fifth-grade ELLs, native speakers of Haitian-Creole from two speech communities, were given the same set of mathematics items in the standard English and standard Haitian-Creole dialects (Sample 1) or in the standard and local dialects of Haitian-Creole (Samples 2 and 3). The largest measurement error observed was produced by the interaction of student, item, and code. Our results indicate that the reliability and dependability of ELL achievement measures is affected by two facts that operate in combination: Each test item poses a unique set of linguistic challenges and each student has a unique set of linguistic strengths and weaknesses. This sensitivity to language appears to take place at the level of dialect. Also, students from different speech communities within the same broad linguistic group may differ considerably in the number of items needed to obtain dependable measures of their academic achievement. Whether students are tested in English or in their first language, dialect variation needs to be considered if language as a source of measurement error is to be effectively addressed.

Keywords: English-language learners, generalizability theory, testing, linguistic minorities

In this article, we address the need for practices in the testing of English-language learners (ELLs) that are supported by concepts from the field of linguistics (see August & Hakuta, 1997; LaCelle-Peterson & Rivera, 1994; Lee, 2002; Lee & Fradd, 1998; Solano-Flores & Nelson-Barber, 2001; Solano-Flores & Trumbull, 2003). To date, approaches to handling linguistic diversity in testing have been limited to the use of classical theory and item response theory (e.g., Ercikan, 1998; Hambleton, Swaminathan, & Rogers, 1991; Van de Vijver & Tanzer, 1998). We propose the use of generalizability (G) theory (Brennan, 1992, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) as a tool for examining the validity of achievement measures for ELLs.

We contend that G theory allows development of testing models that are more sensitive to evidence that bilingual individuals' skills may vary considerably across languages (Bialystok, 1997; Solano-Flores & Li, 2006; Valdés & Figueroa, 1994). Existing approaches to ELL testing are mainly based on comparing ELLs' and mainstream students' scores (e.g., Abedi, Lord, Hofstetter, & Baker, 2001; Shepard, Taylor, & Betebenner, 1998), or comparing scores between ELLs who do and do not receive testing accommodations intended to reduce the adverse impact of language (e.g., Abedi, Hofstetter, & Lord, 2004; Butler & Stevens, 1997). By contrast, the use of G theory enables examining of the dependability of scores obtained by ELLs tested across languages or dialects.

G theory allows examination of the amount of score variation due to the main effect and interaction effect of student (the object of measurement) and various facets (sources of measurement error) such as item, rater, occasion, and type of task (e.g., Baxter, Shavelson, Goldman, & Pine, 1992; Gao, Shavelson, & Baxter, 1994; Kane, 1982; Ruiz-Primo & Shavelson, 1996; Webb, Schlackman, & Sugrue, 2000). In this study, we included code as a facet. Thus, we examined the amount of score variation due to code and to its interaction with student and with other facets.[1]

Code refers to "any kind of system that two or more people employ for communication" (Wardhaugh, 2002, p. 87).

Guillermo Solano-Flores is Associate Professor of Bilingual Education and English as a Second Language, University of Colorado at Boulder, School of Education, 249 UCB, Boulder, CO 80309-0249; Guillermo.Solano@colorado.edu. His areas of specialization include educational measurement, testing of English-language learners, test development, and test review methods.

Min Li is Assistant Professor of Educational Psychology at the University of Washington, College of Education, BOX 353600, Seattle, WA 98195; minli@u.washington.edu. Her areas of specialization include educational measurement and classroom assessment.

We use this term to refer to either a language or a dialect. A *dialect* is a variety of a language that is distinguished from other varieties of the same language by its pronunciation, grammar, vocabulary, discourse conventions, and other linguistic features (Crystal, 1997).

Although *dialect* is frequently used to mean a substandard or corrupted version of a language, in fact everyone speaks a dialect (Preston, 1993). There may be many dialects of the same language. Standard English is one among many English dialects (Wardhaugh, 2002). Important aspects in which dialect versions of a test differ have to do with the frequency of words and certain syntactical forms. For example, "How much money does Harry need to pay for two popsicles that cost 35 cents each?" could reflect language usage in a given speech community better than "How much money does Harry need to pay if he buys two 35-cent popsicles?" Although the two forms of the same question look similar, forms such as *35-cent popsicles* may be more difficult for students from certain speech communities (see Solano-Flores & Trumbull, 2003).

In this study, we tested ELLs with the same set of items in two codes and examined the reliability and dependability of measures of their academic achievement. We also determined the optimum number of items needed to obtain dependable scores with different ELL testing models.

Linguists (e.g., Cummins, 1999; Krashen, 1996) warn that classifications of ELLs based on programs (e.g., full immersion, bilingual) may be flawed because programs vary considerably in type and fidelity of implementation and because their effectiveness is shaped by multiple contextual factors. Thus, we also examined how the dependability of measures of academic achievement may vary across samples of students who are classified within the same category of "English language learners," yet may have different migration histories and different sets of formal and instructional experiences with their first and second languages.

Since the investigation involved only one broad linguistic group (native speakers of Haitian-Creole), establishing a set of principles for testing all linguistic groups of ELLs is beyond the scope of this article. Rather, we focus on illustrating how new, promising approaches to linguistic diversity in testing can be devised by using the methods and reasonings from G theory.

Methods

Participants

Participants in this study were native speakers of Haitian-Creole, one of the fastest-growing languages spoken in the country (U.S. Census Bureau, 2000). A language spoken by about 6 million people in Haiti, the Dominican Republic, Canada, Puerto Rico, and the United States, Haitian-Creole is considered to be a Romance language, but a unique kind of Romance language: Although it is lexically based on French, it has considerable morphological and syntactic influences from West African languages. Haitian-Creole is a highly codified grammar whose sentences are characterized by explicit rules (Valdman, 1970). Haitian-Creole is not a dialect of French, "but rather a completely independent language, about as closely related to French as modern Italian is to Latin" (Hall, 1953, cited in UCLA Language Materials Project, 2003).

The study's participants were 170 grade 4 and grade 5 students who had been identified as ELLs by their school districts and who were native speakers of Haitian-Creole. These students were enrolled in public schools in City A, Florida, and City B, New York.[2] Florida and New York are the two states with the highest numbers of Haitian-Creole speakers aged 5 years and up in the United States (U.S. Census Bureau, 2000).

Though the three samples consisted of students that would be classified within the category of "English-language learner," they reflected the enormous variation in English proficiency that may exist among individuals within the same given linguistic group and even within the same type of bilingual programs. Sample 1 consisted of 49 students from City A who had been enrolled in bilingual, Home Language Acquisition, or Curriculum Content in Home Language support programs for one and a half to three years. Samples 2 and 3 consisted of 42 students from City A and 79 students from City B, respectively, who had migrated recently from Haiti and were being taught in their native language or had been transitioned recently to full immersion, English-only programs.

Six teachers from City A and seven teachers from City B, from the same school districts as the participating students, also took part in the study. These teachers were all born in Haiti, were native speakers of Haitian-Creole, and lived in the same communities as the participating students. Four of the teachers from each site also participated as raters. These eight individuals were not the teachers of the students whose responses they scored.

Test Item Selection and Translation

We assembled a test consisting of 12 open-ended items selected from the set of items included in the NAEP public releases of mathematics items (National Assessment of Educational Progress, 1996, 2000). The sample only included items that did not contain visual information (e.g., illustrations, graphs, or tables).

The items were translated from Standard English into three Haitian-Creole dialects: Standard, A, and B. A well-established test translation company created the standard version. This company hired two professional translators, native speakers of Haitian-Creole who were born and raised in Haiti. The translators were asked to translate the items into a standard version of Haitian-Creole that could be understood by any Haitian-Creole speaker. [3]

The procedure of back translation—in which a test is translated back into the source language and compared to the original version to see if the original meaning is preserved (see Behling & Law, 2000)—was discarded, as there is mounting evidence that it renders inaccurate evaluations of the quality of translations (see Grisay, 2002). Also, we avoided using a scholarly approach such as one based on lexicological studies, because it would be costly and impractical (see Tanzer, 2005). Instead, we used a pragmatic translation approach. We facilitated separate translation sessions with the City A and City B teams. These teams were instructed to translate the items in a way that reflected the dialect spoken in their communities and that could be understood by their students.

All translation decisions were made collectively by each team after a lengthy process of discussion and review.[4] Although we facilitated the translation sessions, we did not participate in any translation decisions made by these teams. Figure 1 shows an item in its original Standard English version and the three dialect versions of Haitian-Creole.

English, Standard	Jill needs to earn $45.00 for a class trip. She earns $2.00 each day on Mondays, Tuesdays, and Wednesdays, and $3.00 each day on Thursdays, Fridays, and Saturdays. She does not work on Sundays. How many weeks will it take her to earn $45.00? Answer: _____
Haitian-Creole, Standard (testing company)	Jill bezwen sanble 45,00$ pou li ale yon kote ak elèv nan klas li. Li touche 2,00$ chak jou pou lendi, madi, ak mèkredi, epi 3,00$ chak jou pou jedi, vandredi, ak samdi. Li pa travay dimanch. Konbyen semenn sa pral pran pou l rive touche 45,00$? Repons: _____
Haitian-Creole, City A (teachers from City A)	Jil bezwen $45.00 pou l ale nan pwomnad klas la. Nan travay li, li touche $2.00 chak lendi, madi, ak mèkredi, epi $3.00 chak jedi, vandredi, ak samdi. Li pa travay dimanch. Konbyen semèn lap pran pou l genyen $45.00? Repons: _____
Haitian-Creole, City B (teachers from City B)	Jill (Jil) bezwen fè $45,00 pou l ale nan yon pwomnad avèk klas li. Li touche $2,00 chak lendi, chak madi ak chak mèkredi. Li touche $3,00 chak jedi, chak vandredi ak chak samdi. Li pa travay lèdimanch. Konbyen semèn l ap pran l pou l fè $45,00? Repons: _____

FIGURE 1. Original Standard English dialect version and three Haitian-Creole dialect translations of the same item. Differences observed in the translations involve spelling (e.g., week: *semèn* and *semenn*), choice of words (e.g., earn: *touche, genyen,* and *fè*), notation conventions (e.g., $*45.00* and *45,00*$), and discourse style (e.g., each Monday, Tuesday, and Wednesday . . . : *chak lendi, madi, ak mèkredi . . .* and *chak lendi, chak madi, ak chak mèkredi . . .*). Although all these are issues in translation, spelling differences are more frequently observed in Haitian-Creole than in other languages, which reflects the fact that Haitian-Creole has different spelling systems (see Schieffelin & Doucet, 1994). However, unlike the spelling systems used in English, all the spelling systems in Haitian-Creole have a more consistent correspondence between sounds and the forms used to represent those sounds (Mason, 2000).

Test Administration

Students from Sample 1 were tested across languages (i.e., in Standard English and Standard Haitian-Creole). Students from Samples 2 and 3 were tested across dialects (i.e., in the standard dialect and their local dialect of Haitian-Creole).[5] Students were not instructed to respond in one code or another.

Students completed two test booklets, administered in two 45-minute sessions, 7 to 10 days apart. The composition of the booklets can be described generically as follows: Items (a) through (f) were in Code 1 in the first booklet and in Code 2 in the second booklet; items (g) through (l) were in Code 2 in the first booklet and in Code 1 in the second booklet. The sequence of items within each booklet was assigned randomly for each student.[6]

After we assembled the test, we realized that a numeral in one item had been translated with words by one of the translation teams—which might affect equivalence across languages. We also found out that we could not provide the kind of small plastic rulers provided by NAEP for students to draw a rectangle in their response to another item. Although we had the students respond to all 12 items in the two codes, these two items were excluded from our nalyses.

Scoring and Scaling

Raters were trained to use the scoring rubrics used by NAEP. These rubrics (in English) were used to score all responses regardless of code. Training and calibration materials included student responses to items administered in Standard English and in the three Haitian-Creole dialects. These responses were randomly selected and were not part of the corpus of responses analyzed in the study. Raters were given booklets with copies of the student responses arranged by item. To control for the effects of fatigue or practice, each rater was randomly assigned a unique sequence of responses in each booklet.

To deal with the fact that the items used in our study had different point scales (ranging from 2 to 5 points), the scores assigned by teachers were rescaled into a scale extending from 0 to 1. Thus, a dichotomous item was scored 0–1.000 whereas, for example, a 4-point scale item was rescaled 0, .333, .667, 1.000. This treatment allowed us to use all items together in our G analyses, as it preserved both the rank ordering of the students and the absolute level of performance.

Data Analyses

As a first analysis, we performed a series of dependent samples *t*-tests to examine the statistical significance of the test score differences across codes. Also, we performed a series of Pearson correlation analyses to examine the extent to which tests in different codes produced similar student rankings.

We performed G theory analyses with *urGENOVA,* a software package for G theory analyses (Brennan, 2001; The University of Iowa Center for Advanced Studies in Measurement and Assessment, 2004).[7] The basic formulas (based on Brennan, 1992; Cronbach et al., 1972; Shavelson & Webb, 1991) used for estimating the variance components and the relative and absolute decision G coefficients are shown in Figure 2.

We performed a series of $s \times r \times i \times c$ G studies to examine the magnitude of the score variation due to the main

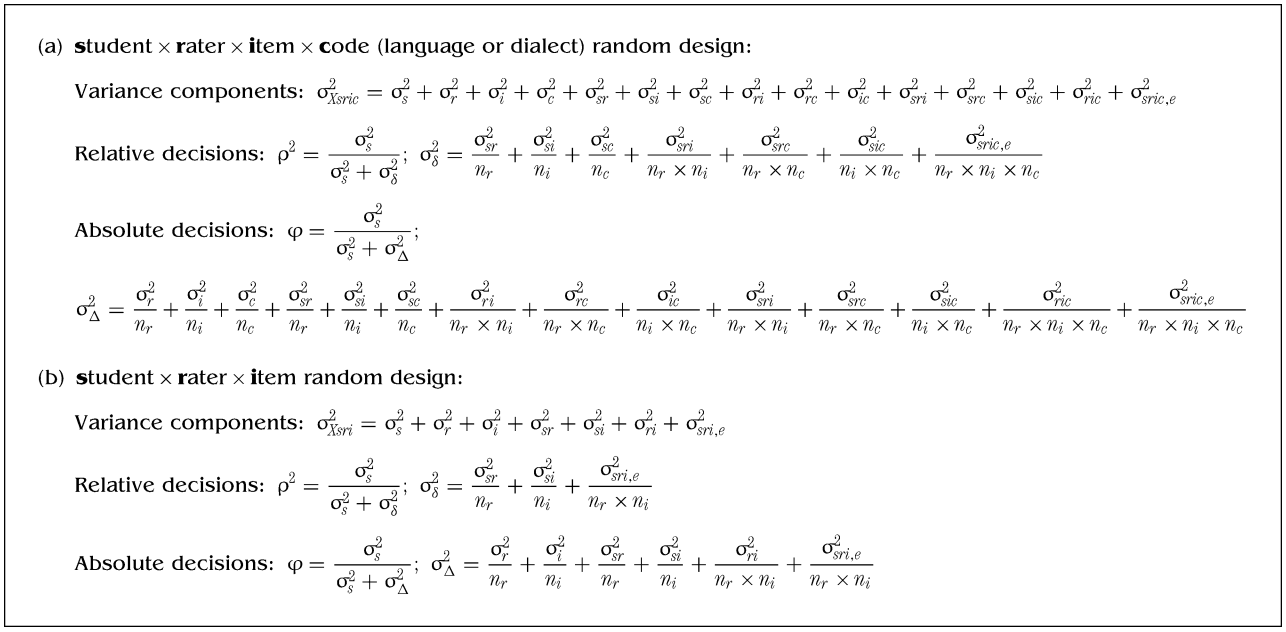

FIGURE 2. Formulas used to calculate variance components and relative and absolute decisions coefficients.

and interaction effect of **s**tudent, **r**ater, **i**tem, and **c**ode (language or dialect). In this design, the total observed score variance (σ^2_{Xsric}) is decomposed into the 15 terms that result from the main and interaction effects of student, rater, item, and code.

Whereas the person score variance (σ^2_s)—estimated from the mean student scores across raters, items, and codes—represents the object of measurement, the remainder of the variance components are estimates of multiple sources of measurement error that result from the sampling variability of rater, item, code, and a combination of these three. The last term (*sric*, *e*), called the error term, cannot be interpreted because the interaction effect of student, rater, item, and code is confounded with error (*e*) due to other, unknown sources of score variation.

Using concepts from G theory, we also examined the generalizability coefficient ρ^2 (equivalent to classical test theory's reliability coefficient), which expresses the reliability of decisions concerning the relative standing of students, and the dependability coefficient φ (a criterion-referenced coefficient), which refers to the absolute level of an individual's knowledge of a given domain. These coefficients express, respectively, how generalizable achievement measures depend on whether they are intended to rank individuals or to index their absolute level of performance (Shavelson & Webb, 1991).

In our analyses, we used random models that assumed that the raters, items, and codes included in the study were samples of a larger universe of raters, items, and codes. When different samples of the same facet are used over replications, the facet has to be regarded as random (Brennan, 1992). This is especially important to understand why we conceptualized code as a random facet: There are multiple forms in which an item can be written in a given language. Each form is a unique combination of lexical and grammatical features that are consistent with the rules of that language; each is a sample of the many possible ways in which that language can be used to write that item (Solano-Flores, 2006). The same reasoning applies to dialect. For example, if several translators were asked to independently translate an item into the standard dialect of a given language, they would produce translations that would not be identical.

Results and Discussion

Mean Score Differences

Table 1 compares the scores obtained by each sample of students on the test in two codes. Statistically significant differences ($\alpha = .05$) were observed for Sample 1. Students from this sample scored higher on items administered in Standard English than in Standard Haitian-Creole. Though students from Samples 2 and 3 performed better on items administered in their local dialect than in the standard dialect of Haitian-Creole, the differences were statistically significant only for Sample 3. Correlations of .701, .817, and .854 between scores obtained in two codes were observed for Samples 1, 2, and 3 respectively, indicating that scores from tests in two codes—especially dialects of the native language—tended to rank students' performance in a roughly similar way.

G Studies: Code (Language or Dialect) as a Source of Measurement Error

We performed a series of random two-facet model ($s \times r \times i$) and three-facet model ($s \times r \times i \times c$) G studies with each student sample. Consistent with results obtained with monolingual populations (Baxter, Shavelson, Goldman, & Pine, 1992; Ruiz-Primo, Baxter, & Shavelson, 1993; Solano-Flores, Jovanovic, Shavelson, & Bachman, 1999), we observed considerable score variation (49% to 61%) due to the $s \times i$ interaction in the two-facet model (Table 2). However, when code was introduced as a facet in the analyses, the relative magnitudes of these effects changed. In the three-facet model, the $s \times i$ interaction effect was reduced to 11% to 23%, whereas the largest score variation was produced by the $s \times i \times c$ interaction (39%, 33%, and 38% for Samples 1, 2, and 3, respectively) (Table 3).

Taken altogether, the results shown in Tables 2 and 3 indicate that, for ELLs, a considerable amount of the score

Table 1. Mean Score Differences on Tests Administered in Two Codes

	Sample: 1 City A ($n = 49$)		Sample 2: City A ($n = 42$)		Sample 3: City B ($n = 79$)	
	Standard Haitian-Creole	Standard English	Local (City A) Haitian-Creole	Standard Haitian-Creole	Local (City B) Haitian-Creole	Standard Haitian-Creole
Mean	.1680	.2997	.1990	.1775	.1809	.1544
SD	.1814	.2518	.2154	.1778	.1730	.1624
Std. error of the mean	.0259	.0360	.0332	.0274	.0198	.0183
p^a	.000		.269		.013	

[a]Paired *t*-test.

variation associated with code is hidden in the $s \times i$ interaction when the two-facet model $s \times r \times i$ is used. That is, the substantial $s \times i$ interaction in the design that excluded code as a facet (see Table 2) was largely due to the variation in the students' responses across codes.

These results indicate that ELL performance is inconsistent across both item and code. In addition to cognitive and content knowledge demands, each item poses a different set of linguistic challenges in each code for each student. A given student performs better in Code X than in Code Y for some items but better in Code Y than in Code X for other items. The results also indicate that the interaction of dialect with student and item is as important as the interaction of language with student and item as a source of measurement error.

Relative and Absolute Interpretation of Scores

We examined the generalizability and dependability of achievement measures and, more specifically, the ρ^2 and φ coefficients obtained by testing ELLs in each of two codes (Table 4).[8] Coefficient φ—the absolute decisions, criterion-referenced coefficient—is especially relevant to interpreting students' test scores in the context of standard-based reform and accountability (Linn, Baker, & Betebenner, 2002).

For students tested across languages (Sample 1), higher G coefficients were obtained when they were tested in Standard English ($\rho^2 = .33$; $\varphi = .31$) than in Standard Haitian-Creole ($\rho^2 = .25$; $\varphi = .23$). For students tested across dialects, the results were inconsistent. For students from Sample 2, higher G coefficients were obtained for scores on items in the local dialect of Haitian-Creole ($\rho^2 = .29$; $\varphi = .28$) than for items in Standard Haitian-Creole ($\rho^2 = .20$; $\varphi = .19$). In contrast, for Sample 3 students, similar G coefficients were observed for scores on items in local Haitian-Creole ($\rho^2 = .21$; $\varphi = .19$) and items in standard Haitian-Creole ($\rho^2 = .21$; $\varphi = .18$).

These results suggest that language as a source of measurement error operates at the level of dialect. Regardless of the language in which ELLs are tested, their performance is sensitive to the dialect of the language in which they are tested. This may be especially true for young ELLs who have had limited opportunities to develop cognitive academic proficiency in their native language (see Cummins, 1984, 1999, 2003; Guerrero, 1997; Hakuta, Butler, & Witt, 2000; Krashen, 1996).

Decision Studies

We performed a series of decision (D) studies (see Shavelson & Webb, 1991) to devise ways in which the considerable magnitude of error due to the interaction effect of code could be minimized. We focused on eight testing models: (a) Monolingual-Standard, (b) Monodialectal-Standard, (c) Monolingual-Native, (d) Monodialectal-Local, (e) Bilingual-Basic, (f) Bidialectal-Basic, (g) Bilingual-Nonequivalent, and (h) Bidialectal-Nonequivalent (Table 5). Each model is defined by whether the test items are administered in one or two codes and, for the tests administered in two codes, by whether two code versions of the same given item are treated as different items.

For each testing model, there is a specific D study design. We calculated relative and absolute error by dividing the estimated variance components by different numbers of items. (Given the small measurement error due to rater observed, the frequency of rater was kept equal to 1.) Then we determined the minimum number of items needed to obtain dependable scores for each testing model.

Table 5 shows the relative decisions and absolute decisions coefficients obtained from the D studies.[9] For students tested across languages (Sample 1), the Monolingual-Standard model rendered the highest G coefficients with the minimum number of items. About 10 items administered in Standard English would be needed to obtain ρ^2 and φ coefficients of .80. In contrast, between 10 and 15 items administered in Standard Haitian-Creole would be needed to obtain ρ^2 and φ coefficients of .80.[10]

For students from Sample 2, tested across their local dialect and the standard dialect of Haitian-Creole, the Monodialectal-Local testing model rendered the highest G coefficients. A little more than 10 items administered in the local Haitian-Creole would be needed to obtain ρ^2 and φ coefficients of .80. By contrast, over 15 items administered in Standard Haitian-Creole would be needed to obtain ρ^2 and φ coefficients of .80.

For students from Sample 3, also tested across their local dialect and the standard dialect of Haitian-Creole, the Monodialectal-Local testing model and the Monodialectal-Standard testing model rendered comparable G coefficients. Slightly less than 20 items administered in either the standard or the local Haitian-Creole dialect would be needed to obtain ρ^2 and φ coefficients of .80.

Note that the differences in the minimum numbers of items needed to obtain dependable scores do not necessarily correspond to the statistical significance of mean score differences observed across codes. For example,

though testing students from Sample 2 in the local and standard dialects of Haitian-Creole did not produce statistically significant mean score differences (see Table 1), fewer items would be needed to obtain dependable scores if they were tested in their local dialect than in the standard Haitian-Creole dialect. By contrast, though testing students from Sample 3 in the local and standard Haitian-Creole dialects produced statistically significant mean score differences, similar G coefficients would be obtained if they were tested in either their local dialect or the standard Haitian-Creole dialect.

In sum, the minimum number of items needed to obtain dependable scores for ELLs may vary depending on both the dialect of the language in which they are tested and the specific characteristics of each group of students.

Summary and Conclusions

We have discussed and illustrated the use of generalizability (G) theory as an

Table 2. Estimated Variance Components and Percentage of Score Variation for Scores on Items Administered in Different Codes: Random *s* × *r* × *i* Model. Relative Error and Absolute Error Indicated

(a) Sample 1. City A, Across Languages

Source of Variability	Standard English			Standard Haitian-Creole		
	N	EVC	%	*n*	EVC	%
Student (*s*)	49	.0533	31	49	.0265	23
Rater (*r*)	4	.0001	0	4	0[b]	0
Item (*i*)	10	.0099	6	10	.0127	11
sr		0[b]	0		0[b]	0
si		.0829	49		.0676	58
ri		.0004	0		.0003	0
sri,e		.0231	14		.0099	8
Rel. error		.1060			.0775	
Abs. error		.1164			.0905	

(b) Sample 2. City A, Across Dialects

Source of Variability	Standard Haitian-Creole			A Haitian-Creole		
	N	EVC	%	*n*	EVC	%
Student (*s*)	42	.0232	19	42	.0375	28
Rater (*r*)	4	0[b]	0	4	0	0
Item (*i*)	10	.0070	6	10	.0064	5
sr		0[b]	0		0[b]	0
si		.0742	61		.0707	52
ri		.0004	0		.0001	0
sri,e		.0167	14		.0204	15
Rel. error		.0909			.0911	
Abs. error		.0983			.0976	

(c) Sample 3. City B, Across Dialects

Source of Variability	Standard Haitian-Creole			B Haitian-Creole		
	N	EVC	%[a]	*n*	EVC	%
Student (*s*)	79	.0214	18	79	.0242	19
Rater (*r*)	4	0	0	4	.0001	0
Item (*i*)	10	.0162	14	10	.0143	11
sr		0	0		.0001	0
si		.0691	59		.0794	61
ri		.0001	0		.0002	0
sri,e		.0095	8		.0113	9
Rel. error		.0786			.0908	
Abs. error		.0949			.1054	

[a]Percentages do not sum up to 100 due to rounding.

[b]Small negative variance components (ranging from −.00027 to −.00001 with a median of −.00007). Compared with the rest of the estimated variance components, these negative values were considered to be negligible and were set to zero, following Brennan's (1992) approach.

Table 3. Estimated Variance Components: *s* × *r* × *i* × *c* Random Model

Source of Variability	Sample 1, City A Std. English, Std. Haitian-Creole			Sample 2, City A Std. Haitian-Creole, A Haitian-Creole			Sample 3, City B Std. Haitian-Creole, B Haitian-Creole		
	n	EVC	%	*n*	EVC	%[a]	*n*	EVC	%[a]
Student (*s*)	49	.0299	20	42	.0277	22	79	.0235	19
Rater (*r*)	4	0[b]	0	4	0[b]	0	4	0	0
Item (*i*)	10	.0097	6	10	.0066	5	10	.0151	12
Code (*c*)	2	.0074	5	2	.0001	0	2	.0003	0
sr		0[b]	0		0[b]	0		0[b]	0
si		.0164	11		.0302	23		.0267	22
sc		.0100	7		.0027	2		0[b]	0
ri		.0004	0		.0005	0		.0002	0
rc		.0001	0		0[b]	0		0[b]	0
ic		.0016	1		.0002	0		.0001	0
sri		0[b]	0		0	0		.0015	1
src		0[b]	0		0[b]	0		.0002	0
sic		.0589	39		.0422	33		.0475	38
ric		0[b]	0		0[b]	0		0[b]	0
sric,e		.0166	11		.0186	14		.0089	7
Rel. error		.1019			.0937			.0848	
Abs. error		.1211			.1011			.1005	
ρ^2		.23			.23			.22	
φ		.20			.22			.19	

[a]Percentages do not sum up to 100 due to rounding.

[b]Small negative variance components (ranging from −.00068 to −.00000 with a median of −.00007). Compared with the rest of the estimated variance components, these negative values were considered to be negligible and were set to zero, following Brennan's (1992) approach.

approach to ELL testing that focuses on code (language or dialect) as a source of measurement error. We have observed that the interaction of student, item, rater, and code is the most important source of error in the measurement of ELL student academic achievement. Our findings particularly underscore the relevance of dialect-level analyses in the testing of ELLs. The score variation due to the interaction of student, item, and dialect can be as large as or larger than the score variation due to the interaction of student, item, and language. Whether ELLs are tested in English or in their native language, the numbers of items needed to obtain dependable scores may vary depending on the dialect of the language in which they are tested.

We also found that groups of students that, according to current testing practices, would be considered as having comparable linguistic proficiencies differed on the minimum number of items needed to obtain dependable scores when they were tested in one dialect or in another. This finding appears to be an effect of the fact that bilingual development is rarely symmetrical; experience and context (e.g., place of origin, migration history, experience with academic language, type of bilingual program, fidelity of implementation of bilingual program) shape the extent to which bilingual individuals are more proficient in either their first or their second language (Bialystok, 2001). Testing ELLs based on broad categories of English proficiency may fail to address important linguistic differences among students from the same broad linguistic group and may affect the dependability of their achievement scores.

Whereas the reasonings and methods described in this article hold in the testing of ELLs, regardless of linguistic group, it is important to note that the basic findings reported here do not seem to be limited to individuals who are native speakers of Haitian-Creole. In a study conducted with native Spanish speakers (Solano-Flores & Li,

Table 4. Relative Decisions (ρ^2) and Absolute Decisions (φ) Coefficients for Scores on Items Administered in Different Codes: *s* × *r* × *i* Random Models

	Sample 1. City A, Across Languages		Sample 2. City A, Across Dialects		Sample 3. City B, Across Dialects	
	Std. English	Std. Haitian-Creole	Std. Haitian-Creole	A Haitian-Creole	Std. Haitian-Creole	B Haitian-Creole
ρ^2	.33	.25	.20	.29	.21	.21
φ	.31	.23	.19	.28	.18	.19

Table 5. Relative Decisions and Absolute Decisions Coefficients Obtained with Different Testing Models and Their Corresponding G Theory Models (Facet Frequencies and D Study Design Shown in Italics)

		Monolingual-Standard[a] *s × r × i*			Monolingual-Native[b] *s × r × i*			Bilingual-Basic[c] *s × r × (i:c)*			Bilingual-Nonequivalent[d] *s × r × i × c*		
		10 Items *r = 1 i = 10 c = 1*	**20 Items** *r = 1 i = 20 c = 1*	**30 Items** *r = 1 i = 30 c = 1*	**10 Items** *r = 1 i = 10 c = 1*	**20 Items** *r = 1 i = 20 c = 1*	**30 Items** *r = 1 i = 30 c = 1*	**10 Items** *r = 1 i:c = 10 c = 2*	**20 Items** *r = 1 i:c = 20 c = 2*	**30 Items** *r = 1 i:c = 30 c = 2*	**10 Items** *r = 1 i = 5 c = 2*	**20 Items** *r = 1 i = 10 c = 2*	**30 Items** *r = 1 i = 15 c = 2*
Sample 1, City A	ρ^2	.83	.91	.94	.77	.87	.91	.68	.76	.79	.65	.72	.77
	φ	.82	.90	.93	.74	.85	.90	.61	.68	.71	.58	.66	.70
		Monodialectal-Standard[e] *s × r × i*			Monodialectal-Local[f] *s × r × i*			Bidialectal-Basic[g] *s × r × (i:c)*			Bidialectal-Nonequivalent[h] *s × r × i × c*		
		10 Items *r = 1 i = 10 c = 1*	**20 Items** *r = 1 i = 20 c = 1*	**30 Items** *r = 1 i = 30 c = 1*	**10 Items** *r = 1 i = 10 c = 1*	**20 Items** *r = 1 i = 20 c = 1*	**30 Items** *r = 1 i = 30 c = 1*	**10 Items** *r = 1 i:c = 10 c = 2*	**20 Items** *r = 1 i:c = 20 c = 2*	**30 Items** *r = 1 i:c = 30 c = 2*	**10 Items** *r = 1 i = 5 c = 2*	**20 Items** *r = 1 i = 10 c = 2*	**30 Items** *r = 1 i = 15 c = 2*
Sample 2, City A	ρ^2	.72	.84	.88	.80	.89	.92	.73	.83	.87	.67	.79	.84
	φ	.70	.83	.88	.79	.89	.92	.71	.82	.86	.65	.78	.83
Sample 3, City B	ρ^2	.73	.84	.89	.73	.84	.89	.73	.85	.89	.67	.80	.86
	φ	.69	.82	.87	.70	.81	.87	.70	.82	.87	.62	.76	.82

[a]All items are administered in Standard English.
[b]All items are administered in the standard dialect of the native language.
[c]Half of the items are administered in Standard English and half are administered in the standard dialect of the native language.
[d]All items are administered in both Standard English and the standard dialect of the native language; the two language versions of the same given item are treated as two different items.
[e]All items are administered in the standard dialect of the native language.
[f]All items are administered in the local dialect of the native language.
[g]Half of the items are administered in the standard dialect and half are administered in the local dialect of the native language.
[h]All items are administered in both the standard dialect and the local dialect of the native language; the two dialect versions of the same given item are treated as two different items.

2006), we also observed that the largest source of score variation was the interaction of student, item, and code, and that the minimum number of items needed to obtain dependable scores varied depending on the dialect of the language in which students were tested. Academic performance instability across languages and across dialects appears likely to be a common occurrence, rather than an isolated phenomenon particular to a specific linguistic group.

Appropriate generalizations of our findings for future research and practice in ELL testing should be based on a careful consideration of the contexts in which testing takes place. For example, though we observed a negligible score variation due to the main and interaction effect of the rater, this does not mean that measurement error due to rater is never an issue in ELL testing. Larger main and interaction effects of this facet might have been observed had we used raters who were not native speakers of the students' native language or who had not been selected carefully. Unfortunately, reports on the testing of ELLs do not provide detailed information about the raters' linguistic qualifications or backgrounds, or the process used to assess the raters' linguistic abilities. The effect of rater as a source of measurement error may have gone unnoticed or may have been underestimated in previous ELL testing research.

The use of G theory in the testing of linguistically diverse populations opens new possibilities for policy and practice in ELL testing. For example, whether a given group of students should be tested in English or their native language could be determined based on testing them in two languages and examining the dependability of their scores. Also, different groups of ELLs within the same broad linguistic group might need to be tested with different numbers of items in order to produce scores of comparable dependability.

Of course, this set of possibilities brings with it a set of challenges and questions. Further research is needed to streamline the process of test translation or local adaptation so that trustworthy dialect versions of the same test can be efficiently developed and to devise assessment systems in which different groups of students can be given different numbers of items for the same test. Moreover, given the observed relevance of dialect, further research is needed to determine whether, when testing students only in English, more dependable scores can be obtained if they are tested in their local dialect of English than in Standard English.

For now, our results show that G theory can be used to devise sound, effective psychometric approaches to testing ELLs that address linguistic diversity in a way that is consistent with current thinking in the field of linguistics.

Notes

This investigation was supported by the National Science Foundation, Grants REC-0126344, REC-0336744, and REC-0450090. We are grateful to Rich Shavelson for his advice

and to Elise Trumbull and Xiaohong Gao for their collegial support. We are indebted to the participant schools and students for welcoming us in their communities and classrooms, and to the teachers, who enthusiastically participated in this project. We wish to acknowledge Melissa Kwon, Ursula Sexton, and Cecilia Speroni for their valuable participation in data collection; and Phil Esra for his comments on the article. We are also grateful to Steve Ferrara and three anonymous reviewers for their insightful comments. The opinions expressed in this report are those of the authors and do not necessarily reflect the position, policy, or endorsement of the funding agency. Nor do the opinions expressed here necessarily reflect those of our colleagues.

[1]In this article, *testing in two codes* means administering the two code versions of each given item at different times. This approach differs from other approaches (e.g., Anderson, Liu, Swierzbin, Thurlow, & Belinski, 2000), in which the test format displays two code versions of the same item side by side.

[2]Letters are used to keep the identities of the participating sites and schools confidential.

[3]Following the company's procedures, one translator translated the test and the other reviewed and commented on the translation; a revised version of the translation was created after any disagreement was discussed and worked out.

[4]The entire process of translation took, on average, about $1\frac{1}{2}$ and 2 hours per item, respectively for City A and City B. First, teachers broke into pairs and translated the item. (Several pair combinations of teachers were used across items, so that each teacher would work together with other teachers in translating at least one item.) Then one of the pairs presented its translation to the other teachers, who proposed changes based on their knowledge of the language spoken by their students and their own experience in translating the item. This group discussion continued until consensus was reached. After completing this process for each item, another review was made and a final version was agreed upon. Teachers were provided with an English-Haitian-Creole-English dictionary and documents released by NAEP that detailed the mathematical knowledge and skills each item intended to address. The translation procedure is discussed in detail by Solano-Flores, Speroni, and Sexton (2005).

[5]In this article, *testing students in two (or across) languages*, means testing them in the standard dialects of those languages. Similarly, *testing students across dialects* means testing them in the local and standard dialects of their native language.

[6]In a previous study (Solano-Flores, Lara, Sexton, & Navarrete, 2001), we used a Latin-square design to compare a reduced number of item sequence blocks in their effects on the performance of Grade 4 ELLs on NAEP science and mathematics items given in two languages. Since no effects were observed in that study, in this study we used a design that addressed the possible effects of fatigue, practice, or maturation more effectively.

[7]Although *urGenova* does not provide standard error of measurement (SEM) estimates for unbalanced designs, model misspecification is discarded, given the negligible magnitude of the negative variance components. In addition, only .29% of the values were missing (39 of 13,600 scores), which eliminates the possibility that the negative values are a reflection of sampling variability due to small sample sizes of the estimated variance components (Brennan, 2001). In a similar study conducted with native Spanish-speaking ELLs and no missing data at all (Solano-Flores & Li, 2006; see the Conclusions section), we found very small SEM for all estimates.

[8]The results reported are based on the scores given by raters from the same sites as the students (i.e., scores given by teachers from City A to students from City A and scores given by teachers from City B to students from City B). In a series of supplementary G studies (available from the authors upon request) we addressed the fact that raters were nested within sites. These nested-design studies revealed that the main effect and interaction effect of rater nested within site were negligible.

[9]The full tables with the estimated variance components are available upon request.

[10]The value of .80 is used here just as a reference value for comparison purposes. Assessment systems may use higher coefficients as criteria for technical quality.

References

Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2001). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16–26.

Anderson, M., Liu, K., Swierzbin, B., Thurlow, M., & Belinski, J. (2000). *Bilingual accommodations for limited English proficient students on statewide reading tests: Phase 2* (Minnesota Rep. No. 31). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved June 10, 2002, from http://education.unm.edu/NCEO/OnlinePubs/MnReport31.html.

August, D., & Hakuta, K. (Eds.) (1997). *Improving schooling for language minority students: A research agenda*. Committee on Developing a Research Agenda on the Education of Limited-English-proficient and Bilingual Students, Board on Children, Youth, and Families, Commission on Behavioral and Social Sciences and Education, National Research Council, Institute of Medicine. Washington, DC: National Academy Press.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement, 29*(1), 1–17.

Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions*. Thousand Oaks, CA: Sage.

Bialystok, E. (1997). Effects of bilingualism and biliteracy on children's emerging concepts of print. *Developmental Psychology, 3*(3), 429–440.

Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition*. Cambridge, UK: Cambridge University Press.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA: The American College Testing Program.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Butler, F. A., & Stevens, R. (1997). Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations. CSE Technical Report 448. National Center for Research on Evaluation, Standards, and Student Testing. University of California, Los Angeles.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Crystal, D. (1997). *The Cambridge encyclopedia of language* (2nd ed.). Cambridge, UK: Cambridge University Press.

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy*. Clevedon, UK: Multilingual Matters.

Cummins, J. (1999). Alternative paradigms in bilingual education research: Does theory have a place? *Educational Researcher, 28*(7), 26–32, 41.

Cummins, J. (2003). BICS and CALP: Origins and rationale for the distinction. In C. B. Paulston & G. R. Tucker (Eds.), *Sociolinguistics: The essential readings*. Oxford, UK: Blackwell.

Ercikan, K. (1998). Translation effects in international assessment. *International Journal of Educational Research, 29,* 543–553.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7,* 323–342.

Grisay, A. (2002). Translation and cultural appropriateness of the test and survey material. In R. Adams & M. Wu (Eds.), *PISA 2000 Technical Report* (pp. 57–70). Paris: Organization for Economic Co-operation and Development.

Guerrero, M. D. (1997). Spanish academic language proficiency: The case of bilingual education teachers in the U.S. *Bilingual Research Journal, 21*(1). Available at: http://www.ncela.gwu.edu/pubs/nabe/brj/v21.htm. Retrieved: April 1, 2004.

Hakuta, K., Butler, Y. B., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* The University of California Linguistic Minority Research Institute Policy Report 2000-1.

Hall, R. A. (1953). Haitian Creole: Grammar, texts, vocabulary. *The American Anthropologist, 55*(2, part 2), Memoire No. 74.

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

Kane, M. T. (1982). A sampling model of validity. *Applied Psychological Measurement, 6,* 125–160.

Krashen, S. D. (1996). *Under attack: The case against bilingual education*. Culver City, CA: Language Education Associates.

LaCelle-Peterson, M. W., & Rivera, C. (1994). Is it real for all kids: A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64*(1), 55–75.

Lee, O. (2002). Promoting scientific inquiry with elementary students from diverse cultures and languages. *Review of Research in Education, 26,* 23–69.

Lee, O., & Fradd, S. H. (1998). Science for all, including students from non-English language backgrounds. *Educational Researcher, 27*(4), 12–21.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3–16.

Mason, M. (2000). Spelling issues for Haitian Creole authoring and translation workflow. *International Journal for Language and Documentation*, *4*(March), pp. 28–30.

National Assessment of Educational Progress. (1996). *Mathematics items public release*. Washington, DC: Author.

National Assessment of Educational Progress. (2000). *Mathematics items public release*. Washington, DC: Author.

Preston, D. R. (Ed.) (1993). *American dialect research*. Philadelphia: John Benjamins.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30,* 41–53.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996, April). *Concept-map based assessment: On possible sources of sampling variability*. Paper presented at the American Educational Research Association annual meeting, New York.

Schieffelin, B. B., & Doucet, R. C. (1994). The "real" Haitian Creole: Ideology, metalinguistics, and orthographic choice. *American Ethnologist, 21*(1), 176–200.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22–27.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shepard, L., Taylor, G., & Betebenner, D. (1998). *Inclusion of limited-English proficient students in Rhode Island's Grade 4 mathematics performance assessment*. CSE Technical Report No. 486. Center for the Study of Evaluation; National Center for Research on Evaluation, Standards, and Student Testing; Graduate School of Education and Information Studies, University of California, Los Angeles.

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English-language learners. *Teachers College Record* (in press).

Solano-Flores, G., Jovanovic, J., Shavelson, R. J., & Bachman, M. (1999). On the development and evaluation of a shell for generating science performance assessments. *International Journal of Science Education, 21*(3), 293–315.

Solano-Flores, G., Lara, J., Sexton, U., & Navarrete, C. (2001). *Testing English language learners: A sampler of student responses to science and mathematics test items*. Washington, DC: Council of Chief State School Officers.

Solano-Flores, G., & Li, M. (2006). Examining the dependability of academic achievement measures for English-language learners: Validity, accommodations, and testing model implications (Under review).

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553–573.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32*(2), 3–13.

Solano-Flores, G., Speroni, C., & Sexton, U. (2005). *Test translation: Advantages and challenges of a socio-linguistic approach*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, April 11–15.

Tanzer, N. K. (2005). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In R. Hambleton, P. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Erlbaum.

The University of Iowa Center for Advanced Studies in Measurement and Assessment (2004). Computer Programs. http://www.education.uiowa.edu/casma/computer_programs.htm.

UCLA Language Materials Project (2003). Haitian Creole profile. http://www.lmp.ucla.edu/profiles/profh01.htm. Retrieved February 18, 2005.

U.S. Census Bureau (2000). Census 2000, Summary File 3, Table PCT10. Internet release data: February 25, 2003.

Valdman, A. (1970). *Basic course in Haitian Creole*. Bloomington: Indiana University Press.

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex.

Van de Vijver, F. J. R., & Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47*(4), 263–279.

Wardhaugh, R. (2002). *An introduction to sociolinguistics* (4th ed.). Oxford, UK: Blackwell.

Webb, N. M., Schlackman, J., & Sugrue, B. (2000). *The dependability and interchangeability of assessment methods in science*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST) technical report 515. Los Angeles: University of California, Los Angeles.