DOCUMENT RESUME

ED 423 254 TM 028 989

AUTHOR	Auchter, Joan E.; Skaggs, Gary; Stansfield, Charles
TITLE	Linking Tests across Two Languages: Focus on the Screening of Biliterate Hispanic U.S. Seniors.
PUB DATE	1998-04-00
NOTE	22p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Bilingual Students; English; Equivalency Tests; *High School Seniors; High Schools; *Hispanic Americans; Item Bias; Sampling; Screening Tests; *Spanish; Tables (Data); *Test Construction; Test Format; *Translation
IDENTIFIERS	*General Educational Development Tests; Hispanic American Students; Linkage

ABSTRACT

A multi-year effort is being made to create a revised Spanish-language version of the Tests of General Educational Development (GED). It is necessary to ensure that the translated, adapted version maintains the same content and performance standards as the original English version. The final linking of the Spanish-language and English versions calls for a design that involves the administration of anchor or common items in the two languages to one biliterate sample, a sample that is equally proficient in both languages. This study evaluated the screening procedure for identifying and selecting graduating high school seniors who are equally literate in Spanish and English. A test that could be used for this purpose was developed based on the fourth GED test, "Interpreting Literature and the Arts," a test that does not rely on prior knowledge of literary works or familiarity with the language of literary analysis. The developed screening test was administered to 500 Hispanic high school seniors in Florida and California. In practical terms, only seniors whose number-correct scores are equal or different by one on the two language halves would be selected as balanced biliterates. In the pilot sample, 36% of the seniors met the stringent GED selection criteria. It was necessary to redo the analysis of differential item bias using only biliterate students after completing the screening. Results of this study appear to validate the screening procedure for identifying and selecting the biliterate students who will be used to link the Spanish-language translations of the GED tests to their corresponding English versions. (Contains 13 tables, 1 figure, and 9 references.) (SLD)

Linking Tests across Two Languages: Focus on the Screening of Biliterate Hispanic U.S. Seniors

Joan E. Auchter, GED Testing Service of the American Council on Education

Gary Skaggs, West Mesa Associates, Inc.

Charles Stansfield, Second Language Testing, Inc.

Abstract

The long term goal of this multi-year effort to create a revised Spanish-language version of the GED Tests is to ensure that the translated, adapted version maintains the same content and performance standard as the original English version. While it is desirable to remove the language barrier to the attainment of a high school diploma for those adults who are literate in Spanish, special care must be taken to link scores in the two languages so that one set of norms and standards can be used. The final linking of the Spanish-language test to the English version calls for a design that involves the administration of anchor or common items in the two languages to one biliterate sample-—a sample that is equally proficient in the two languages. When this procedure is finalized, linking the two versions of the test can proceed. Establishing a valid and practical procedure for the selection of balanced biliterates is the first step towards placing the English and Spanish versions on the same scale. From this group, a sample will be selected which is as similar as possible to the distribution of ability within the 1996 sample of graduating high school seniors that was used to establish the norms and cut scores for the English-language tests. The advantage is that there will be only one set of performance standards (one set of norms) for the GED tests. In addition, the performance of these biliterate students can help isolate differences in the Spanish- and English-language versions of the test instruments that are due to translation. The purpose of this study is to evaluate the screening procedure for identifying and selecting graduating high school seniors who are equally literate in Spanish and English.

Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, April 1998

Linking Tests across Two Languages: Focus on the Screening of Biliterate Hispanic U.S. Seniors

Joan E. Auchter
GED Testing Service

Gary Skaggs
West Mesa Associates, Inc.

Charles Stansfield
Second Language Testing, Inc.

BACKGROUND

Description of the English-language GED Tests. The GED Tests are designed to provide an opportunity for persons who have not graduated from high school to earn a high school level diploma that is recognized by both institutions of higher education and by employers. Administered in all fifty states and the territories, almost 800,000 people take the GED Tests annually. Approximately one in seven high school diplomas issued each year in the U.S. is a GED diploma.

The third and current generation of English-language GED Tests, introduced in 1988 and renormed in 1996, is a five-test battery that requires seven hours and 45 minutes of testing time. The five subtests are: Writing Skills, Social Studies, Science, Interpreting Literature and the Arts, and Mathematics. All five tests are comprised of multiple-choice items. In addition, the Writing Skills Test requires an expository essay.

To allow GED candidates the opportunity to demonstrate achievement comparable to that of high school graduates, the tests are based on two foundations: 1) test content that conforms as closely as possible to the core academic curricula of U.S. high schools, and 2) score scales based on periodic norming of the GED Tests on a nationally representative sample of graduating U.S. high school seniors. This norming process allows the passing standards for the GED Tests to be referenced to the actual performance of those who graduate via the traditional route. The passing standard for the GED Test is set to be somewhat higher than that for graduation from high school. With the 1997 initiation of a higher minimum passing score requirement, over one third of graduating high school seniors would not pass the GED Tests.

Description of the Spanish-language GED Tests. In addition to the core English-language tests, there is a Spanish-language version of the tests, also introduced in 1988. The Spanish-language GED Tests were originally developed to provide adults in Puerto Rico who had not completed high school an opportunity to earn a GED diploma comparable to the diploma awarded by high schools in Puerto Rico. (Spanish is the primary language of instruction in Puerto Rican High Schools.) As a result, the content of the Spanish-language GED Tests reflects the typical high school curriculum in Puerto

Rico. The tests were normed using only on graduating high school seniors in Puerto Rico.

Because of the large number of Spanish-speaking adults in the U.S., many states began offering the Spanish-language GED as an accommodation for candidates with limited English proficiency. Currently, the Spanish-language GED Tests are taken more often in the continental United States than in Puerto Rico. In 1997, approximately 27,00 tests were administered in the mainland U.S., while about 15,000 were administered in Puerto Rico (GEDTS, 1998).

STATEMENT OF THE PROBLEM

The Spanish-language GED Tests do well what they were developed and normed to do: provide an opportunity for adults in Puerto Rico to earn a GED diploma comparable to the diplomas awarded by high schools in Puerto Rico. The use of the Spanish-language tests outside of Puerto Rico has been criticized because some states offer the same high school level credential, regardless of the particular language version of the GED Tests taken. This use may be considered inappropriate because the content of the two language versions varies, the tests are normed on different populations, and the score scales are not linked. Thus, it is possible that different levels of ability are required to obtain the same GED score, and therefore, the credential. There is no evidence to validate the use of the Spanish-language version in the U.S.

As a result of these concerns, the governing boards of the GED Testing Service required the GED Testing Service to develop a new Spanish-language version of the GED Tests that would maintain the English-language content and passing standards and could be used as an accommodation to Spanish-speaking adults.

METHODS

Translatability Study and Feasibility Panels. To determine if the goals of this project are obtainable, the GED Testing Service first conducted a series of preliminary analyses to determine the translatability of the GED Tests, then convened a series of psychometric and linguistic feasibility panels to advise on technical and translation issues. Complete analyses of the translatability study are included in the GED *Direct Translation Feasibility Study,* (Colberg, 1993*).* Psychometric deliberations are presented in the document *Linking the English-language and Spanish-language Versions of the Tests of General Educational Development: Psychometric Feasibility Study,* (Sireci, 1994.*)* Linguistic deliberations are summarized in the *Development of Revised Spanish-language Versions of the Test of General Educational Develpment: Linguistic Feasibility Study* (Auchter, 1996).

Translation Process. The results of these efforts were reported last year at NCME's annual meeting (Auchter & Stansfield, 1997). The paper reports that the GED Tests could be effectively translated. In addition, the paper outlines a rigorous nine-step translation process, that if strictly adhered to, could reduce the likelihood of introducing bias factors that can lead to differences in performance across the translations. This model is based on the Combined Feasibility Panel report and the *Guidelines for Adapting Educational and Psychological Tests* from the International Test Commission (ITC, largely summarized by Hambleton, 1994). All items on three operational forms of four of the five translated subtests (Science, Social Studies, Interpreting Literature and the Arts) are direct translations of the English-language items and potentially can be considered as anchor items in the linking study. Almost half of the Writing Skills Test items are direct translations and potentially can be considered as anchor items. About 20% of the items required that two or more distractors be modified, and another 20% of the translated stems resulted in changes that reflect such Spanish only categories as accent and other diacritic marks. The remaining 10% of the items required new stimulus sentences and options.

Establishing Empirical Links. In addition, the Auchter and Stansfield paper describes a procedure for establishing empirical links between the Spanish- and English-language versions. At the core of this linking design is a procedure for selecting subjects for the final linking sample based on their performance on a screening test in both languages. Only students who demonstrate equal ability across the languages will be selected for inclusion.

Demonstrating that the procedure for the selection of balanced biliterates is valid and practical is the first step towards placing the English and Spanish versions on the same scale. In addition, these biliterate students will be further screened to select a final sample which is as similar as possible to the distribution of ability within the 1996 sample of graduating high school seniors that was used to establish the norms and cut scores for the English-language tests. When a common scale is achieved, the norms for the English language tests will be applicable to both language versions. The advantage is that there will be only one set of performance standards (one set of norms) for the GED tests.

The purpose of this study is to evaluate the validity and feasibility of the proposed screening procedure for identifying and selecting graduating high school seniors who are equally literate in Spanish and English.

Technical Issues Related to Adapted Test Equivalence. It is important to note here that, although we will be using IRT equating methodology, we are not strictly equating Spanish- and English-language tests. According to the framework described by Linn (1993), this linking would best be described as calibration. That is, we will empirically place the Spanish-language tests on the same scale as the U.S. English-language tests. When the common scale is achieved, the norms for the English-language tests will be applicable to both language versions. The advantage is that there will be only one set of performance standards (one set of norms) for the GED Tests. However, due to differences in language, we cannot say that Spanish and English tests are parallel forms of the same construct and that it would be a matter of indifference to an examinee which form he/she takes. Thus the language versions are linked and not equated.

There are two research designs which account for most of the linkings between cross-language test forms. We will describe these designs briefly. They are more completely described by Hambleton (1994) and Sireci (1997). In the first design, separate language versions of the test are given separately to samples of monolingual examinees. Items that appear to be functioning equally in the two languages form an anchor test which can be used to link the separate language forms. This design has two major difficulties. One, there can be a confounding between monolingual sample performance differences and test translation differences. Two, the two monolingual samples may be sufficiently different in performance to make linking problematical.

The second design is the use of bilingual sample. The goal of this design is to eliminate the confounding of the monolingual groups design and isolate what are strictly translation differences. However, the use of this design also has two major drawbacks. First, most bilingual individuals are not equally proficient in both languages, that is *biliterate*, and this can bias the linking. Second, this sample would not be representative enough of either monolingual population to permit a valid linking.

GEDTS' final goal is to directly translate existing tests from the source language—English, to the target language—Spanish, and maintain the English-language content and standards for both language versions. The purpose of this study is to evaluate a procedure for selecting a biliterate sample to link the two language versions of the GED Tests. If we can identify a subsample of biliterate examinees from a bilingual sample, then we can address the first concern of the bilingual design. We will have a sample, of examinees who are equally literate in both languages. The performance of these biliterate students can help isolate differences in the Spanish- and English-language versions of the test instruments that are due to translation. Further, if the biliterate subsample spans a great enough range of ability, with special attention at the cut scores, we can select a final sample as similar as possible to the distribution of ability with the 1996 sample of graduating high school seniors that was used to establish the norms and cut scores for the English-language tests and thus a valid linking. This representative sample would address the second concern of the bilingual sample design, that the sample would not be representative enough of either monolingual population by adequately reproducing the English-language norm sample. The analyses below will address these two bilingual design issues.

SCREENING FOR BILITERATE HISAPNIC U.S. GRADUATING HIGH SCHOOL SENIORS

Development of a biliterate screening test. A test that could be used to screen biliterate seniors was based on the fourth GED Test: Interpreting Literature and the Arts. This passage-based test is essentially a measure of the ability to comprehend and analyze literary selections, and to apply interpretations to new contexts. Items do not rely on prior knowledge of literary works or familiarity with the language of literary analysis or criticism. For the purpose of removing language as a barrier, this test would be a good indicator of biliteracy among adult candidates.

To construct the screening test, six passages were selected from an operational English-language GED Test 4. Two forms of the screening test were created, each using the same six passages and related items. Each form contains two parts, one in English and one in Spanish, having 17 items in one part and 18 items in the other. The language Part 1 and Part 2 were presented in a counterbalanced order in the two forms of the screening tests. Table 1 below shows the type of passage, language, and proportion correct (p-values) for the set of items associated with each passage. The p-values came from the 1996 standardization study of graduating high school seniors (in English). The two parts of three passages each were selected so that they were equal in difficulty, at least in the standardization sample. The overall p-value for Part 1 (passages 1, 5, 6) was .710, while the p-value for Part 2 (passages 2, 3, 4) was .714.

Table 1
Biliterate Screening Test Design

		Language				
	Passage	**Form Y**	**Form Z**	**Genre**	**No. of Items**	**P-value**
Part 1	1	English	Spanish	Fiction	6	.78
p-value	5	English	Spanish	Commentary	6	.65
.710	6	English	Spanish	Poetry	5	.69
Part 2	2	Spanish	English	Drama	6	.75
p-value	3	Spanish	English	Fiction	7	.73
.714	4	Spanish	English	Non-fiction	5	.69

Procedure for Determining Biliteracy. The first task is to equate Part 2 proportion-correct raw scores (P2) to Part 1 scores proportion-correct raw (P1). Although the two parts are nearly equal in difficulty, they are not identical forms and thus need to be equated. To accomplish this, we use the basic equations for linear equating:

$$P2^* = A(P2) + B,$$

where P2* is the P1 equivalent of a P2 raw score, and A and B are the slope and intercept of the equation and are defined as follows:

$$A = \frac{S_{P1}}{S_{P2}}$$

and $B = \overline{P1} - A\left(\overline{P2}\right)$

Based on the 1996 Standardization, A=.951 and B=.024.

The next step is to establish a confidence interval around the following difference score on the two half-length tests:

$DIFF = P1 - P2^{*}$

According to classical measurement theory, a 95 percent confidence interval around a true DIFF score of zero can be expressed as: $\pm(1.96)SEM_{DIFF}$,

where $SEM_{DIFF} = SEM_{P1-P2\bullet}$

and $SEM_{P1-P2\bullet} = \sqrt{SEM^2_{P1} + SEM^2_{P2\bullet} - 2r_{P1,P2\bullet}\left(SEM_{P1}\right)\left(SEM_{P2\bullet}\right)}$

According to the 1996 Standardization data, $SEM_{DIFF} = .0583$, and the confidence interval for determining equal proficiency in English and Spanish would be $\pm.114$. In practical terms, this interval means that we would accept candidates whose English and Spanish number-correct raw scores were equal or differed by one. We understand that this requirement is stringent, but selected this parameter because of the length of the subtests.

Pilot Study Sample. This paper reports a pilot study conducted in May 1977 during which the screening instrument was administered to 500 Hispanic graduating high school seniors in California and Florida. The schools were given direction to select for participation only students they believed to be biliterate—those having equal ability to read and write Spanish and English. Some schools selected students who had exited ESL classes. The sample includes 281 students of primarily Central American derivation (Mexico and San Salvador) representing 8 schools in California and 79 students of primarily Caribbean origins (Cuba and Puerto Rico) representing 5 schools in Florida. This sample should isolate most translation issues related to dialect, region, or cultural issues.

During March through May of 1998, the screening instrument will be administered to approximately 2,000 biliterate high school seniors in California, Texas, Florida, New York and Illinois. These states were selected because they have the greatest number of Hispanic students. The screening instrument will be administered in concert with the Spanish-language adapted translations of the English Test and a second test (spiral of Science, Social Studies, Interpreting Literature and the Arts and Mathematics.) The original design was to administer the screening test first. Subjects would then be selected for the final linking sample using a confidence interval of +.114. In practical terms, only seniors whose number-correct raw scores are equal or different by one on the two language halves would be selected as balance biliterates. For ease of administration in the secondary schools, testing will only occur once. Rather than administering only the screening instrument to the bilingual students, then administering the translated versions to those selected as biliterates, the full sample of students will also take two other Spanish-language tests. Total test time is up to four hours. Data analyses for linking purposes will only include biliterate students whose performance was equal on both halves of the screening instrument. From this subsample, the final sample will be selected which is as similar as possible to the distribution of ability within the 1996 sample of graduating seniors that was used to establish the norms and cut scores for the English-language test.

Data Analyses. The analyses were designed to assess directly the main difficulties of using a bilingual sample for linking: 1) is it possible to identify biliterate students, and 2) if so, is this sample representative enough of their respective monolingual populations to provide a valid linking? For this studies purpose, the second difficulty is modified to address representation of the 1996 sample of graduating seniors that was used to establish the norms and cut scores for the English-language test. To accomplish this, we undertook three sets of analyses.

Item equivalency. The first task in evaluating the screening test is to determine if the items are functioning the same way in each language. For this task, we evaluated the entire bilingual sample. Using BILOG (Mislevy & Bock, 19), we estimated item difficulties using the Rasch model for each item separately in each language. Although a number of IRT models are available, we used the Rasch model in this study due to the sample size constraints. Using a bivariate plot of the item difficulties, we could identify any items which behaved differently relative to the other items.

Results of the screening. Once we had a set of items which we felt functioned equally in both languages, we then carried out the screening to examine how well it worked in identifying a biliterate sample. Additionally, we examined differences in performance of the biliterate students across the two languages on the two halves of the test, and also examined the difference in performance between the biliterate and bilingual populations.

Representativeness of the biliterate sample. Once biliterate examinees were identified, we looked at how representative this sample is of the norming sample of U.S. graduating high school seniors.

RESULTS AND DISCUSSION

Item Equivalency. Using the entire bilingual sample, we analyzed the screening test items to determine if they functioned the same way in both languages. Table 2 below shows proportion correct for each part of the screening test in each language. A comparison of the far right column of rows one and two shows that the entire sample of bilingual students performed better in English than they did in Spanish, .55 compared with .66. A comparison of English-language performance in rows two and three indicates that the average proportion correct for all bilinguals was .66, compared to .71 for the U.S. standardization sample. The English-language performance for this sample was therefore somewhat lower, but surprisingly close to that of English-speaking high school seniors.

While the second row indicates that the two parts of the test were about equally difficult in English, row one indicates that in Spanish, Part 2 is easier than Part 1. By comparing the Spanish- and English-language performance in the column representing Part one, there is a significant difference in proportion correct across the two languages, .52 and .65 respectively. This difference, which is limited to the Spanish translation, suggests that some items in the screening test are probably not functioning the same way in the two languages.

Table 2
Mean Proportion Correct Comparing Screening Test Parts in Two Languages*
With 1996 Standardization Sample of Graduating U.S. Seniors

		Part 1 (Passages 1,5,6)	Part 2 (Passages 2,3,4)	Total
Spanish	Prop. Correct	.52 (.21)	.58 (.20)	.55 (.20)
English	Prop. Correct	.65 (.23)	.66 (.20)	.66 (.21)
English 1996 Norms	Prop. Correct	.710()	.714()	.71()

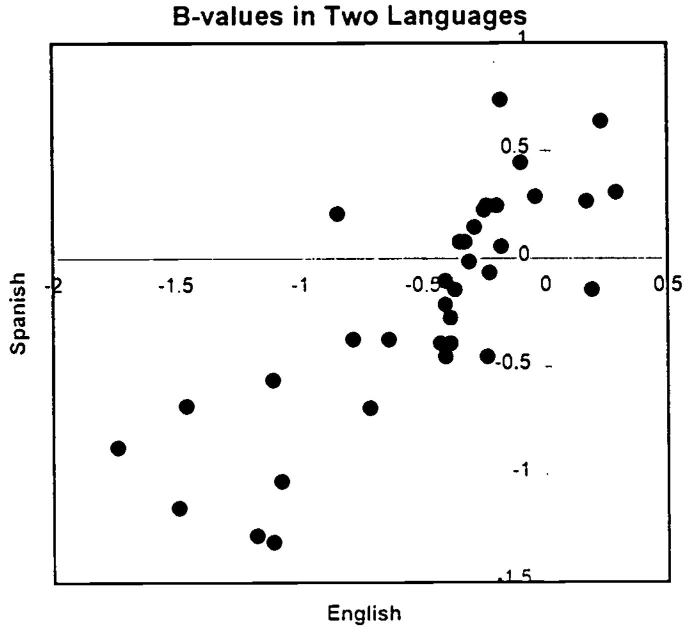
* Standard deviations are in parentheses.

To explore this possibility, we estimated Rasch model item difficulties using the BILOG (Mislevy & Bock, 1989). Although a number of IRT models are available, we felt that sample size limited our analyses here to a one-parameter model. When studying operational data from the Spring 1998 study, we hope to obtain enough data to explore alternative models. Also, due to the biliterate sample size limitation, this estimation is based on the entire bilingual population. If the evaluation were conducted using only the biliterate sample, the correlations might be significantly different.

We estimated separately in each language item difficulties and then placed them on a bivariate scatterplot. This plot is shown in Figure 1. The correlation between the two sets of item difficulties was .80. In equating parallel forms of tests, correlations above .90 are usually found. The figure below represents the same items, in two languages, following a very rigorous translation process. Despite that process, a number of items appear to be functioning differently (i.e. showing differential item functioning, DIF) in the two languages. Unfortunately, the item difficulties were estimated on samples of just over 150. Therefore, the lower than expected correlation is also due to some degree on weak estimation.

However, we attempted to screen out those items with the most differential functioning. To do this, we formed a 95 percent confidence interval around the difference between the two item difficulties based on their standard errors. This interval was typically about ±.50. By this criterion, two items were excluded from the screening test for this study. One of these items, number 14 was a part of passage 3, Part 2. The p-values for this item were .52 in English and .27 in Spanish. The other item, number 27 was a part of passage 5, Part I. The p-values for this item were .82 in English and .43 in Spanish. By excluding them, the correlation between the item difficulties increased to .85.

Figure 1



After excluding the two items, the mean proportions correct from Table 2 were recalculated. These are shown in Table 3 below. Excluding the two high DIF items had the effect of closing the gap between the mean proportions correct in the two parts in Spanish. As a screening test in this study, the remaining analyses were based on the screening test without the two high DIF items. The differences in column one in proportion correct between the Spanish- and English-language versions of Part I are still of concern and will be further addressed with the biliterate sample.

Table 3
Mean Proportion Correct for Bilingual Screening Test Parts in Two Languages With 1996 Standardization Sample of Graduating U.S. Seniors with Two High DIF Items Excluded

		Part 1 (Passages 1,5,6)	Part 2 (Passages 2,3,4)	Total
Spanish	**Prop. Correct**	.53	.57	.55
English	**Prop. Correct**	.68	.67	.67
English 1996 Norms	**Prop. Correct**	.71	.71	.71

Results of the Screening. Table 4 below shows the results of the screening test. Overall, 36 percent of the sample, which was previously identified as bilingual, was selected as biliterate. There was considerable difference between the two forms on the selection rate. The highest selection rate came from students for whom we could not identify the form they had taken (students bubbled in a form letter on the answer sheet).

Table 4
Percent of Sample Selected as Biliterate

	N	N accepted as biliterate	% accepted as biliterate
Form Y	131	54	41
Form Z	121	32	26
no form	45	20	44
Total	291	106	36

In trying to explain these results, we wondered what the impact of deleting the two high DIF items was on the selection rate. If all items had been used, then we would have selected 53 of 131 students on Form Y but 45 of 121 students on Form X, which would have made the selection rates of the two forms more equal. Deleting these two items had a greater impact on Form Z. Since the only difference in Form Y and Form Z is the counterbalancing of languages, this difference is a cause for further investigation.

Table 5 shows the mean proportions correct for the two parts in each language among the selected biliterate subsample. The third column confirms that overall this group scored equally in the two languages. However, there was a difference between the parts. While the bilingual sample scored about the same on English Parts I and Parts 2 (.65 and .66), the biliterate sample scored significantly higher on Part 1, .66 and .59 respectively. Additionally, the biliterate subsample total English performance of .63 is lower that the .66 of the bilingual sample. This difference may be a result of the biliterate inclusion limitation to Hispanic seniors whose number-correct raw scores are equal or different by one on the two language halves. Conversely, this biliterate subsample scored higher in Spanish on Part 2. Noting that the criterion for biliteracy was a difference of .11, these differences of .08 were somewhat less but still a cause for further investigation.

Table 5
Mean Proportion Correct for Screening Test Parts in Two Languages for Biliterate Sample (with Two High DIF Items Excluded) With 1996 Standardization Sample of Graduating U.S. Seniors

		Part 1 (Passages 1,5,6)	Part 2 (Passages 2,3,4)	Total
Spanish	Prop. Correct	.58	.67	.64
English	Prop. Correct	.66	.59	.63
English 1996 Norms	Prop. Correct	.71	.71	.71

To drill down and explore these concerns, we examined the screening test by passage and compared the mean proportion correct for each passage from the U.S. norming group, the bilingual sample, and the biliterate subsamples. These results are shown in Table 6. A passage by passage comparison of performance in English of the bilingual and biliterate samples shows the biliterate groups' performance slightly higher on 5 of the 6 passages, with the greatest difference by the biliterate group on passage 5. The same comparison in Spanish presents somewhat different and mixed results. On passage 5, the two groups perform the same, but with the lowest score of the six passages; performance on this passage also differs the greatest from the U.S. English p value. Additionally, on passages 1, 6, and 4 the biliterate group outperformed the bilingual group, while on passages 2 and 3 the bilingual group outperformed the biliterate group.

The greatest difference occurs across languages within a subgroup. Within the bilingual sample, there was a significant difference in four of six passages, including all three passages in Part 1. We note that the results shown previously in Table 3 compared the three passages in Part 1 with the other three passages in Part 2 in the same language. Of particular interest is the relatively low scores of passages 5 and 6 in Spanish. These passages are based on commentary and poetry.. Since these passages appeared at the end of Form Z, it is difficult to determine if there is a translation effect or a speededness/fatigue effect. An examination of item omit rates indicates that there were higher omit rates for these two prompts overall, but there was no difference in the omit rates between the forms.

Among the biliterate sample, only passages 2 and 5 differed in difficulty between the two languages. The difference in passage 2 was marginally significant, but the difference in passage 5 was larger.

Table 6
Mean Proportion Correct of U.S. Norming, Bilingual, and Biliterate Groups

			U.S.	Bilingual				Biliterate			
	Pass	**Genre**	**P value**	**Form Y**	**P Value**	**Form Z**	**P Value**	**Form Y**	**P Value**	**Form Z**	**P Value**
Part 1	1	Fiction	.78	English	.71*	Spanish	.63	English	.72	Spanish	.71
	5	Commentary	.65	English	.61*	Spanish	.49	English	.67*	Spanish	.49
	6	Poetry	.69	English	.59*	Spanish	.45	English	.60	Spanish	.53
Part 2	2	Drama	.75	Spanish	.70	English	.74	Spanish	.65*	English	.75
	3	Fiction	.73	Spanish	.59	English	.63	Spanish	.56	English	.64
	4	Non-fiction	.69	Spanish	.50*	English	.61	Spanish	.52	English	.59

* p<.05 for difference between language versions within bilingual or biliterate group

Because performance on passage 5 performed most aberrantly. we examined the item difficulties for passage 5 among the biliterate subsample. These results are shown below in Table 7. For all of these items except item 26, the items are significantly easier in English than Spanish. Clearly, with a longer test, we might have excluded the entire passage from the selection process. On the other hand, this is relatively small set of data.

Table 7
P-values of Items from Passage 5 (Commentary)

Item	English P	Spanish P
25	.72	.63
26	.57	.59
27*	.91	.69
28	.63	.38
29	.69	.47
30	.76	.38

• item excluded from selection of biliterates

We had three independent external professional translators answer the items associated with passage 5 to see if they could identify a reason to expect the identified differences. Both translators identified slight miss-translations in items 28 and 30. These miss-translations occurred in the alternatives and introduced potential double-keys. In item 28, the word choice in the correct response did not convey equal strength of meaning as was conveyed in the English-language response; therefore, a second alternative became as correct as the original keyed response. The correct response, "triumphant breakthrough" was translated as "triumphant discovery". In item 30, the keyed response was "cause for celebration" which remained the same; however, an alternative worded "intense suspense" was translated to read "intense astonishment" which became a stronger response that the original key. Since these items are operational, we cannot include the entire items in the paper. Because these miss-translations, isolated by the performance of biliterate seniors, make the items different items in English and Spanish, we decided to exclude these items from the screening process and reanalyze the data. Part I now includes 14 items, while Part 2 includes 17. We recognize that Part 1 is only 14 items long, which is shorter that we would like, but decided to move forward with the rescreening to see if we could identify a biliterate subsample.

Evaluation with items 28 and 30 eliminated. After excluding items 28 and 30, the mean proportions correct from Table 3 were recalculated and are shown in Table 8 below. Excluding items 28 and 30 had the effect of reducing the difference in column one in proportion correct between the Spanish- and the English-language versions from .15 to .10. The difference is now within the .11 criterion for biliteracy.

Table 8
Mean Proportion Correct for Bilingual Screening Test Parts in Two Languages
With 1996 Standardization Sample of Graduating U.S. Seniors
with Two High DIF Items and Items 28 and 30Excluded

		Part 1 (Passages 1,5,6)	Part 2 (Passages 2,3,4)	Total
Spanish	Prop. Correct	.55	.57	.57
English	Prop. Correct	.65	.67	.66
English 1996 Norms	Prop. Correct	.71	.71	.71

Results of the Screening with Items 28 and 30 Deleted. Table 9 shows the results of the screening test with items 28 and 30 deleted. The overall 36 percent of the sample selected as biliterate remained the same. However, the difference in selection rate between the two forms was almost eliminated. Deleting the two items had the greatest impact on Form Z.

Table 9
Percent of Sample Selected as Biliterate

	N	N accepted as biliterate	% accepted as biliterate
Form Y	131	48	37
Form Z	121	43	36
no form	45	15	33
Total	291	106	36

Table 10, excluding items 28 and 30, shows a significant difference from Table 5 in the mean proportions correct for the two parts in each language among the selected biliterate subsample. Removing items 28 and 30 eliminated the differences between the parts as shown in Table 5, and further affirmation of the biliterate sample.

Table 10
Mean Proportion Correct for Screening Test Parts in Two Languages
for Biliterate Sample (with Two High DIF Items and Items 28 and 30 Excluded)
With 1996 Standardization Sample of Graduating U.S. Seniors

		Part 1 (Passages 1,5,6)	Part 2 (Passages 2,3,4)	Total
Spanish	Prop. Correct	.63	.64	.63
English	Prop. Correct	.63	.64	.64
English 1996 Norms	Prop. Correct	.71	.71	.71

After eliminating items 28 and 30, we reexamined the parts of screening test by passage and compared the mean proportion correct for each passage from the U.S. norming group, the bilingual sample, and the biliterate subsample. These results are shown in Table 11. The performance of the bilingual sample remained the same.

Among the biliterate sample, exclusion of items 28 and 30 made a significant difference in performance. None of the six passages differed significantly in difficulty between English- and Spanish-language versions, lending evidence to support our selection of biliterates.

Table 11
Mean Proportion Correct of U.S. Norming, Bilingual, and Biliterate Groups

			U.S.	Bilingual				Biliterate			
	Pass	Genre	P value	Form Y	P Value	Form Z	P Value	Form Y	P Value	Form Z	P Value
Part 1	1	Fiction	.78	English	.71*	Spanish	.63	English	.72	Spanish	.71
	5	Commentary	.65	English	.61*	Spanish	.49	English	.58	Spanish	.63
	6	Poetry	.69	English	.59*	Spanish	.45	English	.56	Spanish	.53
Part 2	2	Drama	.75	Spanish	.70	English	.74	Spanish	.71	English	.75
	3	Fiction	.73	Spanish	.59	English	.63	Spanish	.65	English	.64
	4	Non-fiction	.69	Spanish	.50*	English	.61	Spanish	.54	English	.59

* p<.05 for difference between language versions within bilingual or biliterate group

Representativeness of the biliterate sample. Among the biliterate sample, the mean percent correct was 60 on the entire screening test. Individual raw scores ranged from 10 to 97 percent correct. This means that the biliterate subsample is fairly representative of the parent bilingual sample. The more important question is, how representative is the biliterate subsample of the English U.S. standardization sample?

To answer that question in this study, we compared percentile ranks at ten-percent intervals. From the operational GED Test 4 form from which the screening test was derived, we determined percents correct corresponding to percentile ranks at ten-percent intervals. These percentile ranks are from the 1996 norming study and represent the distribution of English U.S. graduating high school seniors. Then, using the biliterate sample's responses to the screening, we calculated the same percentile ranks within that sample and compared them to the norming sample. These results are shown in Table 12.

Table 12
Comparison of Percents Correct between Biliterate Sample and U.S. Norming Sample at Selected Percentile Ranks

Percentile Rank	U.S. Norming Sample	Biliterate Sample
90	96	89
80	93	76
70	88	71
60	80	61
50	75	58
40	70	52
30	63	47
20	50	43
10	35	31
Mean P	72	60

These results clearly show that the biliterate sample's performance is below that of the U.S. norming sample, but there is at least a wide range of scores within the biliterate sample. The primary issue here is whether the biliterate sample is close enough in performance to the U.S. norming sample to permit a valid linking of Spanish-language tests to occur.

A valid linking in the context of the GED Tests applies to the passing standards that are used. GED Test scores are placed on a scale that ranges from 20 to 80, with a mean of 50 and standard deviation of 10. The GED Testing Service sets minimum passing standards on each of its tests, but states are free to raise those standards. Generally, the passing scale score for a single test is either 40, 45, or 50. We looked at the biliterate sample's distribution relative to those cutpoints. Those results are presented in Table 13 below.

Table 13
Comparison of Percentile Ranks between Biliterate Sample and U.S. Norming Sample at Selected Percentile Ranks

		Percentile Ranks	
Scale Score	Pct. Correct	U.S. Norming Sample	Biliterate Sample
50	80	57	78
45	67	34	66
40	48	16	35

Although the biliterate sample is not representative of the U.S. norming sample, nearly half of the biliterate (43 percent), would be predicted to score between the high and low cutpoints on the GED Test 4.

The use of IRT methodology to conduct the linking offers some degree of invariance. Perfect representation of the two monolingual populations is not required as long as there is a sufficient range of ability in the biliterate sample. We therefore hope that there is enough of a range of scores in the biliterate sample as well as a concentration of scores near the passing scores to permit a valid linking.

SUMMARY

The following two major drawbacks have been identified for current research designs using a bilingual sample to link between cross-language test forms:

- most bilingual individuals are not equally proficient in both languages which can cause linking bias, and
- this bilingual sample would not be representative enough of either monolingual population to permit a valid linking

By summarizing the results of the pilot study conducted to evaluate the screening procedure for identifying and selecting students who are biliterate, that is, can read and write equally well in two languages, we can make the following observations:

- By using a screening test containing two parts of equal difficulty, one part in English and one part in Spanish, it is possible:
  - to select and identify individuals who have equal ability in two languages. Even with the shortened tests halves, 36% of the seniors in this sample met the stringent GED selection criteria of +.11—their raw scores were equal or different by one on the two language halves. Further, none of the six passages differed significantly in difficulty between the English- and the Spanish-language version, lending evidence to support our selection of biliterates.
  - to select a biliterate sample that is representative of the 1996 English-language U.S. standardization sample. While the biliterate sample's performance in English is below that of the U.S. norming sample, there is a wide enough range of scores, as well as a concentration of nearly half (43%) of the biliterate sample between the high and low cut points on the GED standard scale to allow a valid linking.
- It is essential to redo the DIF analysis using only the biliterate students after completing the screening. This second DIF analyisis is how the two items with mis-translation errors were identified. DIF analysis now has another place to work; comparing one language with another. Identification of items using DIF analysis is not unusual. It was interesting to drill down to the passage and then the item level to identify a difference in performance by the biliterate group across the two languages on the same passage. With the elimination of the two items identified as having two potential keys caused by translation word selection in distractors, the biliterate group performance across passages, language halves, and forms balanced.

- It is essential to follow stringent translation procedures to create valid translations from one language to another. Even following the nine step procedure (Hambleton 1994) and (Auchter & Stansfield, 1997), two mis-translations occurred that had the potential of invalidating the screening procedure had not the professional translators isolated the error. It would be beneficial to add a final step to the translation process. In addition to requiring the principal translators to verify keys, a third person who is a native speaker and who has not worked with the tests during the translation process should take the test and identify keys.

- The biliterate selection criterion, a difference of .11 or less on the two halves may be conservative; a wider difference in part scores may be acceptable. The strict criterion in this stud y was selected to compensate in some degree for the shortness of test length, and therefore, lowered reliability of the two test parts. We also wanted to determine the percent selected under a strict criterion to guide the determination of sample size for the Spring 1998 operational study.

- Finding school systems with biliterate students who are willing to commit 4 hours of testing time is a challenge. To secure participation, GEDTS offered a $500 scholarship to the two top scoring students in each state, as well as school reports comparing senior's ability in both languages, and a school report comparing seniors' performance on the GED Tests with that of a national sample of graduating seniors. In addition to the traditional letters requesting participation, to obtain the projected sample of 1,500 to 2,000 seniors, two staff persons have spent two months calling schools, and a "Western Union" mailing was sent to school principals. The biliterate requirement could limit the number of language translations that can be linked to the original standard score scale.

- It would be helpful to have an external validation of biliteracy. The results of this limited sample led to modifications to the Spring 1998 study. Because biliterate students appear to perform better in English than in Spanish, the following questions were added to the answer sheet:
    - In which language did you take your academic courses—English or Spanish? If Spanish, which specific course titles?
    - In which language would you prefer your academic courses—English or Spanish?
    - Which half of the test was easiest to read—English or Spanish?

While it is desirable to provide an accommodation by removing the language barrier, care must be taken when linking scores in two language versions so that one set of norms can be used. The results of this pilot study appear to validate the screening procedure for identifying and selecting biliterate students who will be used to link Spanish-language translations of the GED Tests to their corresponding English versions.

References

Auchter, J.E. (1998). *Who Took The GED? GED 1997 Statistical Report.* Washington, DC: The American Council on Education.

Auchter, J.E. & Stansfield, C.W. (1996). *Development of the revised Spanish-language versions of the Tests of General Educational Development: Linguistic feasibility study.* Washington, DC: GED Testing Service.

Auchter, J.E. & Stansfield, C. W. (1997). *Linking tests across languages: Focus on the translation and adaptation process.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Colberg, M. (1993). *Direct Translation Feasibility Study.* Washington, DC: GED Testing Service.

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

Linn, R.L. (1993). Linking results of different assessments. *Applied Measurement in Education, 6*, 83-102.

Mislevy, R.J. & Bock, R.D. (1989). *BILOG3: Item analysis and test scoring with binary logistic models* (computer program). Mooresville, IN: Scientific Software.

Sireci, S.G. (1994). *Linking theEnglish-language and Spanish-language versions of the Tests of General Educational Development: Recommendations of the GED-STEP Psychometric Feasibility Panel.* Washington, DC: GED Testing Service.

Sireci, S.G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16(1),* 12-19.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM028989

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Linking Tests across Two Languages: Focus on the Screening of Biliterate Hispanic U.S. Seniors

Author(s): Joan E. Auchter, Gary Skaggs, Charles Stansfield

Corporate Source: American Council on Education, West Mesa Associates, Inc., Second Language Testing, Inc.

Publication Date: April, 1998

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Joan E. Auchter

Printed Name/Position/Title: Joan E. Auchter, Exec. Director

Organization/Address: GED Testing Service, American Council on Education, One Dupont Circle
Suite 250, Wash., DC 20036

Telephone: 202-939-9490

FAX: 202-775-8578

E-Mail Address:

Date: 6/23/98

(over)