The impact of linguistic complexity of NAEP test items on 8th-grade students' performance:

Jamal Abedi, Project Director
Advance Research & Data Analyses Center

U.S. Department of Education
National Center for Education Statistics
Grant R999B60012

Advance Research & Data Analyses Center
Santa Monica, CA 90405
(310) 826-2536

The work reported herein was supported under the National Center for Education Statistics Grant No. R999B60012 as administered by the U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the National Center for Education Statistics or the U.S. Department of Education.

TABLE OF CONTENTS

The impact of linguistic complexity of NAEP test items on 8th-grade students' performance:

Jamal Abedi, Fereshteh Hejri and Carol Lord

Abstract

This study examined the degree of impact of students' language background on their academic performance by analyzing the linguistic dimensions and variables which can confound content assessment. The goal of this study was, therefore, to begin to identify linguistic features in NAEP items that may affect the performance of students with language backgrounds other than standard English. Mathematics, science, and geography were selected for this study since content knowledge (not language capability) is the intended target of assessment.

This study used the 1992 main assessment data in mathematics, the 1992 long-term trend data in mathematics and science, and the 1994 NAEP main assessment data in geography, and was conducted in two parts. In the first part of the study, we examined the linguistic complexity of the NAEP test items in the area of math, science, and geography. For evaluating the linguistic complexity of test items, we developed a linguistic rubric based on the literature and the characteristics of items used in the 1992 and 1994 NAEP main and long-term trend assessments. Based on the linguistic rubric, judges assigned ratings of linguistic complexity to the NAEP items. Judges were trained and the inter-judgment consistency was very high. First, linguistic ratings were assigned to the items by trained linguistic experts; then, students' performance in items with different levels of

linguistic complexity was compared across groups of students formed based on their background variables including the language background variables.

The focus of this study was on the language background variables. That is, the main purpose of this study was to find out if language minority students react differently to (perform higher on) the linguistically non-complex items. However, we used other important background variables which have been shown to have impact on students' achievement in school.

In the first part of the study, we compared students' performance in math, science, and geography using the analysis of variance approach. The independent variables in our ANOVA models were background variables including the language background variables. The results of the analyses of variance indicated that language background variables (such as speaking a language other than English at home, or reading English books and newspapers regularly) had substantial impact on students' performance in math, science, and geography.

We also used multiple regression techniques to examine the level of impact of the language background variables on students' performance as compared with the impact of other background variables. The results of multiple regression analyses indicated that in many cases the language background variables had greater impacts on students' performance than many other background characteristics.

In a series of multiple regression models, we included the items' linguistic complexity ratings on the 12 linguistic features which were used in this study along with other item characteristics such as measures of items' length. We also included some subject-level background in addition to the item-level data. A substantial increase in the R^2 was observed due to

addition of the items' linguistic features. A statistical test for significance of the impact of the linguistic features of the test items produced results which were well beyond the .01 nominal level. The results of multiple regression analyses using items' linguistic features indicated that these features are strongly associated with students' performance in the subject area such as math, science, and geography, where the language presumably should not have major impact.

In the second part of the study, we performed a differential item functioning (DIF) analysis on the NAEP test items using language-related variables as grouping variable. We found several items in each subject area (math, science, and geography) that significantly differentiated language minority students from native speakers of English.

Perspective

There is growing concern over the equity of NAEP assessment due to the possible impact of students' language background on their performance. Since NAEP is moving more toward open-ended and extended open-ended items, the issue has become more pronounced. It is hypothesized that language barriers may cause students to score lower in academic subject areas for which language presumably should not be a factor. Non-native English speaking students may obtain lower scores on math and science tests, for example, simply because of problems in communicating or understanding the language of test items or assessment instruments. That is, the students' language background and their performance may be confounded. This study investigated the impact of students' language

background on their performance by examining the effects of the linguistic characteristics of test items on students' performance on those tests.

To make the NAEP assessment more equitable and fair for students with different language backgrounds, the variables related to the students' language background must be recognized at the time of test construction. This study analyzed NAEP data to show how the language background variables and the language of test items could affect students' performance.

The study was conducted in two separate parts. In the first part, items were rated based on their linguistic characteristics, and the performance of students in different subgroups was compared. In the second part, a differential item functioning (DIF) approach was used to examine the possible differential performance across subgroups of students defined on the basis of language background as well as other background variables.

Part One

Comparing test items with different level of linguistic complexity across subgroups of students

In this part, the math test items used in the 1992 main assessment, the math and science items used in the 1992 long-term trend assessment, and the geography items used in the 1994 main assessment were studied for linguistic complexity. Multiple-choice and open-ended items were analyzed separately.

Assigning linguistic complexity ratings to test items

Multiple-choice items were rated on their linguistic complexity by judges using the linguistic rubric developed in this study. For each student, weighted composite scores were created based on the linguistic complexity of the items for each of the three subject areas (i.e., math, science, and

geography). These composite scores were compared across the subgroups of students which were formed based on their background variables including language background. Following is a brief discussion of the approach.

Construction of the linguistic rubric

A rubric was constructed for judging the linguistic complexity of the multiple-choice items. This rubric was developed based on the linguistic characteristics of the NAEP items and based on the literature, including the Language Background Study by Abedi, Lord, and Plummer (1997). The criteria used in the rubric for the Language Background Study were revised; some categories were dropped and others were added.

The criteria for assessing linguistic complexity of the items which were partly developed in the language background studies (Abedi, Lord, and Plummer, 1997) were presented to a group of four linguistic experts to judge the comprehensiveness and the content validity of the criteria. Their suggestions were incorporated into the final version of the rubric. For some features in the rubric, where judges must determine the degree or amount of some characteristic, a Likert-type response mode was adopted. An example of such features is "Familiarity/frequency of non-math vocabulary". For this feature, "very familiar/very frequent" to "non-familiar/rare" categories were used in a five-point Likert scale. For other features where number of occurrences of a feature is the object of judgment, judges were asked to determine the number of occurrences; an example of such features was the number of verbs in the passive voice.

Table 1 lists the 12 features which were used for assessing linguistic complexity of the test items.

Training Judges

A training session was conducted for the judges by the lead linguistic expert. After an introduction to the study, the linguistic rubric was discussed and instruction on how to categorize the sample items was given. The judges were asked to use the linguistic rubric to assign linguistic complexity ratings to a sample of test items. The sample items consisted of some of the 1992 released NAEP math items. The judges were then asked to discuss the rating discrepancies among themselves in reference to the guidelines in the linguistic rubric; the rubric was clarified to provide more explicit guidance to raters. This discussion, supervised by the lead linguistic expert, helped to reach the maximum level of rater/judgment consistency.

Interrater reliability or consistency between judges

Science long-term trend and math long-term trend items were rated by three linguistic experts using the linguistic complexity rubric prepared specifically for this study. Ratings were assigned by each judge on each of the 12 linguistic complexity features. Judges' levels of consistency were obtained by computing the following interrater reliability indices using the ITRS system (Abedi, 1994): (a) percent of agreement, (b) product-moment correlation coefficient, (c) intraclass correlation, (d) Cohen's kappa, and (e) Cronbach's alpha. Table 1 presents a brief description of these features.

Table 1. Linguistic features used in examining the linguistic complexity of the items.

Feature	Description
1	Type of items
2	Participial modifiers
3	Passive voice
4	Modal verbs
5	Medial relatives
6	Final relatives
7	Initial adverbials

8	Medial adverbials
9	Final adverbials
10	Complements
11	Non-preferred subjects
12	Ordinary vocabulary

These statistics were obtained separately for each of the 12 features. Table I1 summarizes the results of interrater reliability analyses for the math long-term trend items. As the data in Table I1 indicates, the interrater-reliability indices for most of the linguistic features are extremely high, and in some cases they are indicative of a perfect interrater reliability. Percent of agreement for most of the features is very high. For the first 8 features, the percent of exact agreement is over 93%, and for those features the percent of agreement within one point range is 100.0 (a perfect agreement). The average P.M. correlation coefficients, however, range from zero to 0.94. For several of the linguistic features, the percent of agreement is near perfect whereas the P.M. correlation is very low or near zero. The low P.M. correlation for some of the features is due to lack of variability of the rating. In many cases, because of the lack of a particular linguistic feature in an item, raters assigned ratings of zero to the items for that feature. Having ratings of mostly zeros for some of the features caused the percent of agreement to increase sharply, but due to lack of variability, the P.M. correlation approached near zero.

The problem of lack of variability affects alpha coefficients too. For some of the features, the alpha coefficient is zero while percent of agreement is near 1. For example, for feature 5, percent of exact agreement is 98.4, but P.M. correlation and alpha is .01. Kappa also is very low for some of the features. This is due to controlling for the chance agreement.

Table I1. Interrater reliability indices for linguistic features using math long-term trend items

Linguistic Features	Pc% of Agreement	Agreement Within 1pnt	P. M Correlation			Kappa		Alpha
			Min	Max	Aver	Coeff	Z	
Feature 1	96.9	100.0	0.91	1.00	0.94	0.94	10.8	0.98
Feature 2	93.7	100.0	0.40	0.86	0.62	0.61	3.1	0.83
Feature 3	93.7	100.0	0.85	0.90	0.87	0.76	5.1	0.93
Feature 4	98.4	100.0	0.66	1.00	0.77	0.77	2.4	0.91
Feature 5	98.4	100.0	0.01	0.02	0.01	0.01	0.0	0.01
Feature 6	93.7	100.0	0.42	0.72	0.60	0.58	2.7	0.81
Feature 7	96.9	100.0	0.57	0.91	0.70	0.68	2.5	0.87
Feature 8	96.1	100.0	0.02	0.04	0.02	0.01	0.0	0.02
Feature 9	82.7	100.0	0.11	0.83	0.47	0.48	3.8	0.75
Feature 10	81.1	99.2	0.02	0.64	0.29	0.21	1.4	0.48
Feature 11	73.2	94.5	0.60	0.92	0.72	0.67	13.5	0.88
Feature 12	70.1	81.1	0.06	0.07	0.04	0.38	6.12	0.04

G-Coefficient = 0.81

A two-facet (8 X 3) analysis of variance repeated measure design was used to estimate the generalizability of the features and raters. In this design, facet A represents linguistic features with 8 levels and facet B represents raters with 3 levels. As the data in Table I1 indicate, a generalizability coefficient of 0.81 was obtained which indicates a fair level of generalizability of features and topics. In this analyses, as it is evident from the interrater reliability indices, the raters were highly consistent. The linguistic feature, however, were not so highly related. This was quite expected because the eight different features were supposed to measure different linguistic aspects of the test items.

Table I2 presents the interrater reliability coefficients for the science long-term trend items. The trend of data in Table I2 is very similar to the trend shown in Table I1. Percent of exact agreement between the raters is very high and in most cases it is near perfect agreement. However, in a few cases a small percent of exact agreement was observed. For example, for Feature 2, a percent of exact agreement of 49.4 was obtained. Similarly, for

Features 10 and 11, percentages were low (59.0 and 60.2 respectively). These low percentages are mainly due to minor disagreement between the raters. This point can be verified by the fact that the percent of within one point agreement for all of the features mentioned above is near perfect.

Similar with the results of interrater reliability analyses reported earlier for the math test items, the interrater reliability indices for the science long-term trend data are not very consistent. As discussed earlier, these inconsistencies are mainly due to assumptions underlying the different statistics and conditions under which those statistics operate. For example, a P.M. correlation coefficient (as well as alpha coefficient) between two perfectly related variables may be zero if there is no variability in the scores of those variables. That is exactly what happened with some of the features. All three raters assigned exactly the same ratings to some of the features; thus, there was no variability in the ratings for those features.

Table I2. Interrater reliability indices for linguistic ratings of the science long-term trend items

Linguistic Features	Pc% of Agreement	Agreement Within 1pnt	P. M Correlation			Kappa		Alpha
			Min	Max	Aver	Coeff	Z	
Feature 1	100.0	100.0	0.00	0.00	0.00	0.00	0.0	0.00
Feature 2	49.4	95.2	0.30	0.69	0.55	0.39	6.5	0.77
Feature 3	83.1	97.6	0.92	0.99	0.94	0.82	14.5	0.98
Feature 4	90.4	100.0	0.93	0.97	0.95	0.84	8.7	0.98
Feature 5	97.6	100.0	0.97	1.00	0.98	0.77	2.4	0.99
Feature 6	91.6	100.0	0.91	0.95	0.93	0.66	3.4	0.97
Feature 7	92.8	100.0	0.74	1.00	0.83	0.77	4.8	0.92
Feature 8	98.8	100.0	0.70	1.00	0.80	0.75	1.5	0.90
Feature 9	66.3	86.8	0.48	0.68	0.57	0.50	6.7	0.80
Feature 10	59.0	91.6	0.53	0.86	0.66	0.52	9.0	0.83
Feature 11	60.2	86.8	0.89	0.94	0.91	0.64	19.8	0.97
Feature 12	97.6	100.0	0.01	0.02	0.02	0.01	0.01	0.03

G-Coefficient = 0.83

In sum, the results of our interrater reliability analyses indicated an extremely high interrater consistency between ratings of the linguistic experts for the linguistic features. This is mainly due to the objectivity of the linguistic rubric and the intensive training session for the raters. This high level of interrater reliability permitted us to use fewer ratings in some of our test items. For some of the test items, we used two ratings rather than three.

Background Variables

The language background variables were the main focus of this study. Students were categorized based on their language background and their test scores were compared. In addition to the language-related variables, other background variables which are shown in the literature to have impact on students' academic progress were also used. Students were grouped based on these variables, and their scores were compared. Among the background variables which were used for categorizing students were gender, ethnicity, parents' education and family socioeconomic status. A list of the relevant background variables including the language background variables which were used in this study is provided in Table 1B.

Table 1B. Background variables used in grouping students

Variable Name	Variable Description
DSEX	Gender
DRACE	Race
STOC	Size and Type of Community
HOMEENV	Home Environment, Reading materials
HOMEEN2	Home Environment, Number of Articles in English
TVWATCH	Television Viewing Each Day
LANGHOM	How Often Other Languages Spoken in the Home
B0033011	How Often People in Home Speak Other Languages

B0035011	Mother's Level of Education
B0036011	Father's Level of Education
B0009011	Does Your Family Get a Newspaper Regularly
B0009021	Is There a Dictionary in Your Home
B0009031	Is there an Encyclopedia In Your Home
B0009041	Are There More Than 25 Books In Your Home
B0009051	Does Your Family Get Magazine Regularly
B0039011	How Much Time Each Day Is Spent On Homework
B0048012	What Kind of Math Class Are You In this Year
S0040011	How Many Days of School Missed Last Month

As indicated earlier, the results of the language background study (Abedi et al., 1997) indicated, for example, that students with very low math performance did not benefit greatly from the language simplification of test items, nor did students with the very high math performance. However, students with average math performance benefited the most from the simplification of items. The current study examined this notion further. In this study, all important background variables were used as bases for categorization, and the hypothesis of differential impact of language complexity on these groups of students was investigated.

The literature has indicated that language minority students benefit from items which are less grammatically complex. However, the literature is scarce with respect to studies examining the impact of a set of comprehensive linguistic factors on students' performance. Will students with different ethnic backgrounds perform differently on items of differing degrees of linguistic complexity? Will students at different levels of SES and parents' level of education perform differently according to the linguistic complexity of the test items? This study aims to provide evidence on these important questions and concerns.

Statistical Design

In this study we performed several different analyses: (a) we compared students' test scores across the subgroups which were formed based on the students' background characteristics including their language background variables; (b) we conducted multiple regression analyses predicting students' scores from their background characteristics; (c) we performed regression analyses on the item-level and subject-level data, and (d) we conducted DIF analyses.

(a) **Analysis of Variance**. Analysis of variance was performed using some of the background variables as independent variables and students' scores on math, science, and geography items as dependent variables. For example, students were categorized with respect to speaking a language other than English in the home. Based on this variable, students were grouped into three categories: (a) always speak that language in the home, (b) sometimes speak that language in the home, or (3) rarely speak that language in the home. This created a one-factor analysis of variance design in which language spoken in the home served as an independent variable and test scores as a dependent variable.

We tested the assumption of homogeneity of variance for the subgroups. No significant difference was found between the subgroup variances, therefore, the subgroups variances on the test scores were assumed to be equal. Also, no major deviation from normal distribution was found in any of the test scores. We also obtained the relevant descriptive statistics such as mean, standard deviation, frequency and graph wherever applicable. We will report the results of some of these descriptive analyses whenever appropriate.

It must be indicated however at this point that any analyses done on the NAEP data must take into account the complexity of NAEP data due to the nature of its design, the balanced incomplete block spiraling (see Beaton, Johnson, & Ferris, 1987; Carlson & Jirele, 1992; Zwick, 1987). Should sampling weights be applied and jackknifing be used in computing different statistics and their standard errors? Since we worked with the item-level data and we wanted to have complete data for each subject, we used the booklet and block-level data.

Results

Analysis of Variance

The background variables shown in Table 1B were used as independent variables in a series of analysis of variance models. In each model, one of the background variables was used as the independent variable, and students' score on math, science, and geography was used as the dependent variable.[1] We will discuss the results of analysis of student background variables in this interim report; analysis and discussion of linguistic characteristics of items will be incorporated and included in the final report.

The results of analysis of variance for each of the background variables in Table 1B were reported in a summary ANOVA table. Tables A1 to A54 in Appendix A presents these ANOVA tables. These tables include source of variation, sum of squares, degrees of freedom, mean square, F-ratio, and the probability of a Type-I error rate.

[1] A major concern in these analyses were the instability of the Type I error rate due to performing multiple independent analyses. On the other hand, including more than one independent variable in the same analysis was not possible due to other limitations and problems. Thus, the results of analysis of variance should be interpreted with caution.

To present an overall picture of these analyses, the results of ANOVA on the background variables were further summarized. Table AA1 shows a summary of results of the ANOVA models using the long-term trend math test scores (Section 1) as the dependent variable. The data in this table include the analysis of variance F-ratio, the p-values, and a coefficient of determination (correlation ratio) which indicates the proportion of the variance of the dependent variable (math score in this case) explained by the independent variable.

As the data in Table AA1 indicate, the ANOVA results show significant differences between the subgroup performance for all of the background variables except for gender. The coefficient of determination, however, show different level of impact of the independent variables on the dependent variable. Some of the background variables (independent variables) had substantial impact on the math test scores. For example, race and "kind of math class student took" each explain over 14% of the variance of math test scores. That is, they are very strong predictors of the math test scores. There are some of the language background variables that are also good predictors of the math test scores. For example, based on the data in Table AA1, "reading materials at home" and "number of articles in English" each explain over 9% of the variance of the dependent variables. It must be indicated at this point, however, that these variables may be confounded by the family SES level.

Background variables which are directly related to the students' language background characteristics such as "How often other languages spoken in the home" produced significant results but did not explain a sizable amount of the variance of the dependent variables. The *language spoken in the home* variable explained only 1.4% of the variance of the dependent

variable. However, based on the results of other analyses which we performed on the language background variables, we believe that there must be a stronger relationship between students' language background and their math performance. This lack of a stronger relationship may be partly due to the small number of students who responded to this question and partly due to other technical problems.

Table AA1. A summary of the results of ANOVAs for Math 1 using students' background variables as independent variables

Variable Description	F-Ratio	P-value	η^2
Gender	3.330	0.068	0.0017
Race	161.460	0.000	0.1486
Size and Type of Community	33.474	0.000	0.0947
Home Environment, Reading materials	98.181	0.000	0.0929
Home Environment, Number of Articles in English	91.227	0.000	0.0950
Television Viewing Each Day	81.228	0.000	0.0779
How Often Other Languages Spoken in the Home	14.004	0.000	0.0144
How Often People in Home Speak Other Languages	14.957	0.000	0.0153
Mother's Level of Education	37.975	0.000	0.0732
Father's Level of Education	54.306	0.000	0.1016
Does Your Family Get a Newspaper Regularly	35.328	0.000	0.0183
Is there an Encyclopedia In Your Home	22.501	0.000	0.0118
Does Your Family Get Magazine Regularly	87.591	0.000	0.0449
How Much Television Do You Usually Watch Every Day	30.052	0.000	0.0859
How Much Time Each Day Is Spent On Homework	9.920	0.000	0.0252
What Kind of Math Class Are You In this Year	148.504	0.000	0.1425
How Many Days of School Missed Last Month	9.739	0.000	0.0158

Table AA2 presented results similar to those presented in Table AA1 for the long-term trend math, Section 2. Consistent with the results shown in Table AA1, students performance across all level of all the background variables were significantly different except for gender. In this table, race and type of math class again are strong predictors of the students' math performance. Variables related to the family SES such as parents' level of education also showed to be strong predictors of students performance in

math. Reading materials at home also explained a sizable amount of the variance of the math scores. Language background variables such as "language spoken in the home" had significant impact on students' performance in math but the impact was not as strong as the *type of math classes*, for example.

Table AA2. A summary of the results of ANOVAs for Math 2 using students' background variables as independent variables

Variable Description	F-Ratio	P-value	η^2
Gender	0.072	0.788	0.0001
Race	103.775	.000	0.1002
Size and Type of Community	29.313	.000	0.0820
Home Environment, Reading materials	76.665	.000	0.0726
Home Environment, Number of Articles in English	72.164	.000	0.0685
Television Viewing Each Day	83.710	.000	0.0784
How Often Other Languages Spoken in the Home	11.496	.000	0.0116
How Often People in Home Speak Other Languages	9.554	.000	0.0096
Mother's Level of Education	45.665	.000	0.0851
Father's Level of Education	64.330	.000	0.1161
Does Your Family Get a Newspaper Regularly	45.996	.000	0.0232
Is there an Encyclopedia In Your Home	11.439	.001	0.0059
Does Your Family Get Magazine Regularly	76.748	.000	0.0387
How Much Television Do You Usually Watch Every Day	28.781	.000	0.0808
How Much Time Each Day Is Spent On Homework	11.513	.000	0.0285
What Kind of Math Class Are You In this Year	300.120	.000	0.2480
How Many Days of School Missed Last Month	18.045	.000	0.0282

Similar to the data reported in Tables AA1 and AA2 for the math items, the data in Table AA3 presents the results for math items Section 3. The results of these analyses are very consistent with those reported earlier for the first two sections. All the background variables (including gender) had significant impact on the students' performance in math. The size of impact, however, differ greatly across the independent variables. "Kind of math class" had the strongest impact followed by parents' level of education. The data in Table AA3 also indicate that the language background variables

had impact on students' performance in math, but the size of impact was not as large as for "kind of math class" and parents' education.

Table AA3. A summary of the results of ANOVAs for Math 3 using students' background variables as independent variables

Variable Description	F-Ratio	P-value	h^2
Gender	7.096	0.008	0.0035
Race	129.322	0.000	0.1208
Size and Type of Community	19.203	0.000	0.0545
Home Environment, Reading materials	70.976	0.000	0.0665
Home Environment, Number of Articles in English	70.122	0.000	0.0657
Television Viewing Each Day	74.821	0.000	0.0698
How Often Other Languages Spoken in the Home	20.936	0.000	0.0206
How Often People in Home Speak Other Languages	27.420	0.000	0.0268
Mother's Level of Education	54.205	0.000	0.0982
Father's Level of Education	70.499	0.000	0.1241
Does Your Family Get a Newspaper Regularly	48.294	0.000	0.0241
Is there an Encyclopedia In Your Home	3.316	0.069	0.0017
Does Your Family Get Magazine Regularly	79.316	0.000	0.0396
How Much Television Do You Usually Watch Every Day	30.188	0.000	0.0834
How Much Time Each Day Is Spent On Homework	7.586	0.000	0.0188
What Kind of Math Class Are You In this Year	205.976	0.000	0.1824
How Many Days of School Missed Last Month	8.214	0.000	0.0131

Multiple Regression Analyses Using Item-Level and Subject-Level Data

As discussed earlier, the test items were rated by linguistic experts on 12 different linguistic features. For each item, we also obtained item-length characteristics such as number of words, number of characters, and number of sentences in the question.[1] These item characteristics (the 12 linguistic ratings and the item-length measures) were used as predictors in a multiple regression model in which the students' total test score was the criterion variable. In addition to the item characteristics mentioned above, we

[1] Since the three measures ofthe item-length (numbers of characters, words, and sentences) were highly correlated (0.94 and higher), we used only one of these measure which is the number of characters.

included some of students' background variables as predictors. A multiple regression model was created for each of the subject areas (math, science, and geography).

To see the impact of the linguistic features on the students' performance on the test items, for each subject area we created two regression models. The predictors for the first model included all 12 linguistic features, item-length characteristics, and some subject-level background variables. These particular background variables (subject level data) were selected because they are likely to have an impact on student scores. This model was labeled the *full model*. The second model, called the *restricted model,* included all the predictors except the 12 item-level linguistic features. Any major increase in the R^2 of the full model as compared with that of the restricted model would be attributable to the effect of linguistic complexity measures that were included only in the full model (the 12 features).

Table R1 summarizes the results of the analyses for the restricted model, which includes all the predictors in the full model except the linguistic features. As the data in Table R1 indicates, all the item-level and subject-level variables contributed significantly to the prediction model. However, the percent of the variance of the dependent variable (long-term trend math score in this case) that was explained by the predictors was only 3.7%.

Table R1. Restricted Model, background variables as predictors (no linguistic features included) and math item score as the criterion variable

Variables in the Equation	B	SE B	Beta	T	Sig T
Number of Characters	-0.0007	0.00002	-0.1705	-37.33	0.000

Home Language Env	-0.0138	0.0054	-0.0129	-2.57	0.010
Lang all	-0.0783	0.0097	-0.0406	-8.08	0.000
Rural	0.0240	0.0090	0.0123	2.66	0.008
Disadvantage urban	-0.1071	0.0074	-0.0676	-14.58	0.000
Advantage urban	0.0448	0.0074	0.0282	6.06	0.000
(Constant)	0.7039	0.0065		107.66	0.000

$R = 0.192$, $R^2 = .037$, $F = 295.26$, $P = .000$

Table R2 presents the multiple regression results for the full model which includes the subject-level variables as well as the 12 item-level linguistic feature ratings. As the data in Table R2 show, all the predictors significantly contributed to the prediction. The multiple R for this model was 0.340, which indicates that the predictors in this model explained about 12% of the variance of the dependent variable. That is, an increase of about 8% on the R^2 was observed due to addition of the linguistic features. The difference between the R^2 of the restricted and full models was tested for statistical significance. An F-ratio of 343.52 (p=0.0000) was obtained, which was significant well beyond the .01 nominal level. These results indicated that the linguistic features of the items are associated with a significant and substantial impact on students' performance in the math subject area, where the language presumably should not have major impact. It is possible that the linguistic complexity of an item is confounded with the item's mathematical complexity; further analyses taking into account the level of mathematical complexity of individual items will be included in the final report.

Table R2. Full Model, Linguistic features and background variables as predictors and math item score as the criterion variable

Variables in the Equation	B	SE B	Beta	T	Sig T
Number of Characters	0.0004	0.00005	-0.0894	-6.99	0.000
Linguistic Feature 2	0.1399	0.0040	0.1804	35.35	0.000
Linguistic Feature 3	-0.0630	0.0020	-0.2180	-32.00	0.000
Linguistic Feature 4	-0.0258	0.0083	-0.0258	-3.13	0.002
Linguistic Feature 6	-0.1211	0.0057	-0.1143	-21.22	0.000
Linguistic Feature 7	0.2601	0.0146	0.1041	17.80	0.000
Linguistic Feature 9	0.1101	0.0036	0.2102	31.01	0.000
Linguistic Feature 10	-0.0198	0.0046	-0.0303	-4.29	0.000
Linguistic Feature 11	-0.0052	0.0024	-0.0304	-2.19	0.029
Home Language Env	-0.0138	0.0052	-0.0129	-2.68	0.008
Lang all	-0.0783	0.0093	-0.0406	-8.42	0.000
Rural	0.0240	0.0087	0.0123	2.77	0.006
Disadvantage urban	-0.1071	0.0071	-0.0676	-15.19	0.000
Advantage urban	0.0448	0.0071	0.0282	6.32	0.000
(Constant)	0.6172	0.0072		85.67	0.000

$R = 0.340$, $R^2 = .116$, $F = 424.06$, $P = .000$

Part Two

Differential item functioning (DIF) analyses based on the language background variables

In a large scale national assessment it is very important to have information on student performance on individual items used for the assessment. This information is helpful in scoring and interpreting test items. In NAEP assessments, individual test items are studied carefully and different statistics are obtained at the item level. Among these statistics is an index of differential item functioning (DIF), used to examine any systematic bias that a test item may have toward a particular group of subjects. NAEP has conducted DIF analyses on items using some of the background variables such as age, gender and ethnicity.

Different approaches have been suggested for examining the possibility of differential functioning of dichotomously scored items (see Allen & Donoghue, 1996). Among the approaches for examining DIF in such items, Quasi-chi-square (Scheuneman, 1975, 1979), log-linear (Alderman & Holland, 1981; Loyd, 1984, and Mellenbergh, 1982), Mantel-Haenszel (MH) (Holland and Thayer, 1988), Standardization procedure (Dorans and Kulick, 1983, 1986), SIBTEST (Shealy and Stout, 1993), and logistic regression approach (Spray and Carlson, 1988) could be mentioned. NAEP has frequently used two different approaches, a graphical method, which could be conducted by the modified version of BILOG program (Mislevy and Bock, 1984), and the MH procedure suggested by Holland and Thayer (1988). Some of these approaches use the observed scores as the major conditioning variable and others use estimates of true scores as the major conditioning variable.

To assess the possible differential item functioning between two groups, examinees from the focal group, (i.e., the group to be studied) must be

matched based on their abilities underlying the performance, θ, to a reference group (the group which is used as standard against which the performance of the focal group is compared). This underlying ability, θ, is estimated based on IRT, or sometimes it is estimated based on the total score of the test. In the MH approach, the subjects in the focal group and the reference group are matched based on the total score of the test (the total number of correct responses on the test). Traditionally, the MH approach is applied in testing programs where all examinees receive the same set of items. This is not the case, however, in NAEP testing programs. Because of the complex nature of the NAEP data, the balanced incomplete block spiraling (see Beaton, Johnson, & Ferris, 1987; Carlson & Jirele, 1992; Zwick, 1987), DIF analyses are usually done at the block or booklet level. Zwick and Grima (1991) compared the results done based on block and booklet levels and described advantages and disadvantages of each approach. Their recommendation was to perform DIF analyses at the block level and to use the total number of items correct in the block as the matching criteria (see also, Allen, & Donoghue, 1991). Thus, the focal and the reference groups are matched based on the total items correct on the block level (Holland & Thayer, 1988; Zwick, R. and Grima, A., 1991; Allen and Donoghue, 1991).

In addition to the complexity of matching the focal and reference group in NAEP, the issues of variance estimation and weights in NAEP data are important for analysis and interpretation of DIF statistics. The relevant question in regard to variance estimation and weights is "should sampling weights and jackknifing be used in computing MH D-DIF statistics and their standard errors?" (Zwick & Grima, 1991, p. 6). The suggestion is "..the best procedure for NAEP use is the method that includes weights, but not jackknifing" (Zwick and Grima, 1991, p. 8).

The Mantel-Haenszel procedure has been performed separately for each NAEP item, and based on the results of MH analyses, items have been categorized into three groups: (1) Group A, items which have no evidence of DIF, (2) group B, items which may have some evidence of DIF and (3) group C, test items with considerable evidence of DIF. The C items are usually marked for modification or deletion from the test (see the NAEP 1990 Technical Report, 1992, p. 155; 1992 Technical Report, 1994).

To do a valid MH analyses on items, there must be enough subjects in the focal and reference groups. The minimum number of subjects suggested is 100 subjects in the smaller group; the focal group and the reference group should have a total of at least 500 subjects (Petersen, 1988). This could create a limitation for using many of the background variables because in some levels of some of the NAEP background variables there may not be enough subjects to perform the analysis. Possibly because of this limitation, NAEP performed DIF analyses only based on gender, ethnicity and in some cases by age. The comparisons were made between whites and Hispanics, whites and African Americans, and males and females. In these analyses, white and male subjects were used as reference groups and Hispanics, African Americans and females were used as focal groups.

In this study we performed DIF analyses based on some language background variables. A comparison based on LEP/non-LEP would be informative. We also performed DIF based on other relevant language background variables. For example, a DIF was performed to compare the native English speakers with those who speak a language other than English at home. Since some of the subgroups had small Ns, different categories of some of the variables were combined. We used the existing NAEP student background data to identify student populations who might be expected to

have difficulty with the language of test items. These populations included, for example, students whose first language was not English, and/or whose parents did not speak English in the home. Groups of language minority (LM) students were identified using these variables or combinations of these variables. Using similar variables, complement groups of students (who are not expected to have difficulty with the language of the test items) were identified.

Based on the recommendations presented earlier, we did match the focal and reference groups based on the total items correct in the block. The software we used for this analysis was developed by Rogers and Hembleton (1994).

To examine any possible differential impact of students' language background on the test items, the LANGHOM variable was selected and was used in our DIF analyses. LANGHOM is one of the questions in the NAEP background questionnaire which asks students: *How often other languages spoken in home*? It has 3 response categories: *1. Never, 2. Sometimes, and 3. Always*. For the DIF analyses we used students who responded *sometimes* or *always* to this questions as the focal group and other students in the group were used as the reference group.

The analyses were performed on the 1992 math main assessment items, the 1992 math and science long-term trend items, and the 1994 geography items. We will report the results for the math long-term trend items in this short report. The complete data will be presented in the final report.

Tables D1 through D12 (Appendix B) summarize the results of DIF analyses performed on the 1992 long-term trend math and science items. Table D1 (Appendix B) reports the results of DIF for the long-term trend

math items (Section 1). Performance of the students who indicated that they always spoke other languages in home were compared with the total group for any differential outcome. Of the 37 math items in this section, 6 showed significant DIF. Of the six significant DIF, in 4 cases, the focal group performed lower than reference group and in 2 cases, the reference group performed lower than the focal group.

Similarly, Table D2 (Appendix B) presents the results for those students who indicated that sometimes they spoke other languages in the home. In this analysis, 2 items showed significant DIF. Of these two items, one had higher score for the focal group and one had higher score for the reference group. A comparison of the results of DIF analyses in Tables D1 and D2 suggests that the math test items are more biased toward students who speak less English at home and who may be less proficient in English than the native speaker.

Table D3 (Appendix B) presents the results of DIF for the long-term trend math (Section 2) for students who always spoke other languages in home. In this table, 5 out of 24 items showed evidence of DIF. In 2 of the 5 cases, the evidence favored the reference group and in 3 cases it favored the focal groups. Table D4 presents similar results for students who indicated that sometimes they spoke other languages. None of the items in Table D4 showed significant DIF. Comparing the data reported in Tables D3 and D4 suggest that there were more differences in math items functioning with the group who claim that they always speak other languages in home.

Tables D5 and D6 (Appendix B) presents the results of DIF for the math long-term trend data (Section 3). The results are very similar to those presented earlier. The results for students in the “always” category show more evidence of DIF than for others. In Table D5, six of the 37 items

showed significant DIF. Of those, three favored the reference group and three favored the focal group.

Tables D7 to D12 presents similar results for the science test items. Comparing the results of DIF for science items across the two categories of "always" and "sometimes" indicates that in the "always" category there is more evidence of DIF.

Discussion

We performed different analyses on the NAEP test items to examine the impact of students' language background on their performance. The results strongly and consistently suggest that students' background characteristics, particularly their language background, do impact their performance in math, science, and geography, areas that traditionally have not been linked with students' verbal abilities.

The results of analyses of variance indicated that most of the students' background variables obtained from the NAEP background questionnaire showed significant impact on students' performance in math, science, and geography. Among these, the effects of the language background variables were quite evident.

The results of multiple regression analyses indicated that the linguistic characteristics of the test items were associated with a substantial impact on students' performance in content areas, where students' scores presumably should not be affected by their English language proficiency.

REFERENCES

Abedi, J. (1994). Interrater/Test Reliability System (ITRS). User's Manual. Los Angeles: Advance Research & Data Analyses Center.

Abedi, J., Baker, E. L., & Herl, H. (1995, August). Comparing reliability indices obtained by different approaches for performance assessments. (CSE Technical Report 401). University of California, Los Angeles.

Abedi, J., Lord, C. & Plummer, J.R. (1997). Language background as a variable in NAEP mathematics performance. National Center for Research on Evaluation, Standards, and Student Testing. University of California, Los Angeles/CRESST.

Adams, M. J. (1990). Beginning to read: Thinking and learning about print. Cambridge, MA: MIT Press.

Aiken, L. R. (1971). Verbal factors and mathematics learning: A review of research. Journal for Research in Mathematics Education, 2, 304-13.

Aiken, L. R. (1972). Language factors in learning mathematics. Review of Education Research, 42(3), 359-85.

Alderman, D.L., & Holland, P.W. (1981). Item performance across native language groups on the Test of English as a Foreign Language (TOEFL Research Rep. No. 9; ETS Research Rep. No. 81-16). Princeton, NJ: Educational Testing Service.

Allen, N. L. and Donoghue, J. R. (1991). Applying the Mantel-Haenszel procedure to complex samples of items. Paper presented at the Annual Meeting of the National Council on Measurement in Education. Chicago.

Allen, N. L. and Donoghue, J. R. (1996?). Detecting differential item functioning: Current methods and continuing problems. New Jersey: Educational Testing Service.?

Baker, E. (1991). Issues in policy, assessment, and equity. Washington, DC: United States Department of Education, Office of Bilingual Education and Minority Languages Affairs.

Beaton, A. E., Johnson, E. G., & Ferris, J. J. (1987). The assignment of exercises to students. In A. E. Beaton (Ed.), Implementing the new design: The NAEP 1983-84 technical report (pp. 97-118). Princeton, NJ: Educational Testing Service.

Bormuth, J. R. (1966). Readability: A new approach. Reading Research Quarterly, 1(3), 79-132.

Botel, M., & Granowsky, A. (1974). A formula for measuring syntactic complexity: A directional effort. Elementary English, 1, 513-516.

Byrne, B. M. (1984). The general/academic self-concept nomological network: A review of construct validation research. Review of Educational Research, 54, 427-456.

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

Carpenter, T. P., Corbitt, M. K., Kepner, H. S., Jr., Linquist, M. M., & Reys, R. E. (1980, September). Solving verbal problems: Results and implications from national assessment. Arithmetic Teacher, 28, 8-12.

Carroll, J. B., Davies, P., & Richman, B. (1971). The American Heritage word frequency list. Boston, MA: Houghton Mifflin.

Celce-Murcia, M., & Larsen-Freeman, D. (1983). The grammar book: An ESL/EFL teacher's book. Rowley, MA: Newbury House.

Chadha, N. K. (1989). Causal antecedents of self-concept, locus of control and academic achievement: A path analysis. Psychologia, An International Journal of Psychology in the Orient, 32(4), 265-268.

Cocking, R. R., & Chipman, S. (1988). Conceptual issues related to mathematics achievement of language minority children. In R. R. Cocking & J. P. Mestre (Eds.), Linguistic and cultural influences on learning mathematics, pp. 17-46. Hillsdale, NJ: Erlbaum Associates.

Cohen, J. (1960). A coefficient of agreement for normal scales. Educational and Psychological Measurement, 20, 37-46.

Cohen, J. (1968). Weighted kappa; Nominal scale agreement with provision for scaled disagreement and partial credit. Psychological Bulletin, 70, 213-220.

Collis, G. M. (1985). Kappa, measure of marginal symmetry and intraclass correlations. Educational and Psychological Measurement, 45, 55-62.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. Cognitive Psychology, 20, 405-438.

Cummins, J. (1984). Bilingualism and special education. San Diego, CA: College Hill Press.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. Educational Research Bulletin, 27, 11-20; 28, 37-54.

Davison, D. M., & Schindler, S. E. (1988). Mathematics and the Indian student. In Reyhner, J. (Ed.). Teaching the Indian child: A bilingual/multicultural approach. Billings, MT: Bilingual Education Program.

De Corte, E., Verschaffel, L., & De Win, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. Journal of Educational Psychology, 77(4), 460-470.

Dorans, N.J., & Kulick, E. (1983). Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach (RR-83-9). Princeton, NJ: Educational Testing Service.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76, 378-382.

Fleiss, J. L., Cohen J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 72, 323-327.

Fleiss, J. L., Nee, J. C. M., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. Psychological Bulletin, 86, 974-977.

Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32, 221-233.

Flynn, T. M. (1991). Achievement, self-concept and locus of control in Black pre-kindergarten children. Early Child Development and Care, 74,135-139.

Fry, E. (1977). Fry's readability graph: Clarification, validity, and extension to level 17. Journal of Reading, 21, 242-252.

Ginsburg, H. (1981). The clinical interview in psychological research on mathematical thinking: Aims, rationales, techniques. For the Learning of Mathematics, 1(3), 4-11

Hagborg, W. J. (1991). Group counseling with adolescent special education students: Challenges and useful procedures. Adolescence, 26(103), 557-563.

Harris, A. J., & Jacobson, M. D. (1973-1974). Some comparisons between the Basic Elementary Reading Vocabularies and other word lists. Reading Research Quarterly, 9, 87-109.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), Test validity. pp. 129-145). Hillsdale, NJ: Erlbaum.

Hudson, T. (1983). Correspondences and numerical differences between disjoint sets. Child Development, 54, 84-90.

Hunt, K. W. (1965). Grammatical structures written at three grade levels (Research Report No. 3). Urbana, IL: National Council of Teachers of English.

Hunt, K. W. (1977). Early blooming and late blooming syntactic structures. In C. R. Cooper & L. Odell (Eds.), Evaluating writing: Describing, measuring, judging. Urbana, IL: National Council of Teachers of English.

Jerman, M., & Rees, R. (1972). Predicting the relative difficulty of verbal arithmetic problems. Educational Studies in Mathematics, 4, 306-323.

Johnson, E. G. & Allen, N. L. (1992) The NAEP 1990 Technical Report. Washington DC: National Center for Education Statistic.

Johnson, E.G. & Allen, N.L. (1994). The NAEP 1992 Technical Report. Washington DC: National Center for Education Statistic.

Jones, P. L. (1982). Learning mathematics in a second language: A problem with more and less. Educational Studies in Mathematics, 13, 269-87.

Klare, G. R. (1974). Assessing readability. Reading Research Quarterly, 10, 62-102.

Larsen, S. C., Parker, R. M., & Trenholme, B. (1978). The effects of syntactic complexity upon arithmetic performance. Educational Studies in Mathematics, 21, 83-90.

Lepik, M. (1990). Algebraic word problems: Role of linguistic and structural variables. Educational Studies in Mathematics, 21, 83-90.

Lorge, I. (1939). Predicting reading difficulty of selections for children. Elementary English Review, 16, 229-233.

Loyd, B. (1984, April). Evaluation of log-linear models for detection of item bias: A comparison across samples. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Lyon, M. A., & MacDonald, N. T. (1990). Academic self-concept as a predictor of achievement for a sample of elementary school students. Psychological Reports, 66(3),1135-1142.

Maqsud, M., & Rouhani, S. (1991). Relationships between socioeconomic status, locus of control, self-concept, and academic achievement of Batswana adolescents. Journal of Youth and Adolescence, 20(1), 107-114.

Marsh, H. W., Walker, R., & Debus, R. (1991). Subject specific components of academic self-concept and self-efficacy. Contemporary Educational Psychology, 16(4), 331-345.

McNemara, J. (1966). Bilingualism in primary education. Edinburgh: Edinburgh University Press.

Mestre, J. P. (1988). The role of language comprehension in mathematics and problem solving. In R. R. Cocking & J. P. Mestre (Eds.), Linguistic and cultural influences on learning mathematics (pp. 201-220). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., & Bock R.D. (1984). BILOG: Marginal estimation of item parameters and subject ability under binary logistic models. Chicago: International Educational Services.

Peterson, N. S. (1988). DIF procedure for use in statistical analysis. NJ: Educational Testing Service.

Posner, K. L., Sampson, P. D., Caplan, R. A., Ward, R. J., & Cheney, F. W. (1990). Measuring interrater reliability among multiple raters: An example of methods for nominal data. Statistics in Medicine, 9, 1103-1115.

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.), The development of mathematical thinking (pp. 153-196). New York: Academic Press.

Rogers, J.H., and Hambleton, R.K. (1994). MH: A FPRTRAN 77 program to compute the Mantel-Haenszel statistic for detecting differential item functioning. Educational and Psychological Measurement, 54, 1, 101-104.

Saxe, G. B. (1988). Linking language with mathematics achievement: Problems and prospects. In R. R. Cocking & J. P. Mestre (Eds.), Linguistic and cultural influences on learning mathematics (pp. 47-62). Hillside, NJ: Lawrence Erlbaum Associates.

Scheuneman, J.D. (1975, April). A new method of assessing bias in test items. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 106 359).

Scheuneman, J.D. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Shealy, R.T., & Stout, W.F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometricka, 58, 159-194.

Spache, G. D. (1953). A new readability formula for primary-grade reading materials. Elementary School Journal, 53, 410-413.

Spanos, G., Rhodes, N. C., Dale, T. C., & Crandall, J. (1988). Linguistic features of mathematical problem solving: Insights and applications. In R. R. Cocking & J. P. Mestre (Eds.), Linguistic and Cultural Influences on Learning Mathematics (pp. 221-240). Hillsdale, NJ: Erlbaum Associates.

Spray, J.A., & Carlson, J.E. (1988). Comparison of loglinear and logistic regression model for detecting changes in proportions. (Research Rep. No. 88-3). Iowa City, IA: American College Testing.

Sue, D. W., & Sue, D. (1990). Counseling the culturally different. Theory and practice. New York: John Wiley & Sons.

Travers, K. J. (1988). Opportunity to learn mathematics in the eighth-grade classrooms in the United States: Some findings from the Second International Mathematics Study. In R. R. Cocking & J. P. Mestre (Eds.), Linguistic and cultural influences on learning mathematics, pp. 187-199. Hillsdale, NJ: Erlbaum Associates.

Wang, M. D. (1970). The role of syntactic complexity as a determiner of comprehensibility. Journal of Verbal Learning and Verbal Behavior, 9, 398-404.

Zwick, R, and Grima, A. (1991). Policy for differential items functioning (DIF) analysis in NAEP. NJ: Educational Testing Service.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. Journal of Educational Measurement, 24, 293-308.

Zwick, R., Donoghue, J. R. and Grima, A. (1993). Assessment of differential items functioning for performance tasks. Journal of Educational Measurement, 30, 3, 233-251.

Appendex A

Results of analysis of variance using background variables as independent variables and test scores as dependent variables

Section 1

Table A1. ANOVA Results for Math Scores by Gender

Source of Variation	SS	df	MS	F	P
Gender	452.345	1	452.345	3.330	.068
Within Subjects	261660.74	1926	135.857		
Total	262113.09	1927	136.021		

Table A2 ANOVA Results for Math Scores by RACE

Source of Variation	SS	df	MS	F	P
RACE	37227.487	2	18613.74	161.460	.00
Within Subjects	213160.51	1849	115.284		
Total	250387.99	1851	135.272		

Table A3. ANOVA Results for Math Scores by Size and Type of Community

Source of Variation	SS	df	MS	F	P
Size and Type of Commun	24810.385	6	4135.064	33.474	.00
Within Subjects	237302.70	1921	123.531		
Total	262113.08	1927	136.021		

Table A4. ANOVA Results for Math Scores by Reading Materials

Source of Variation	SS	df	MS	F	P
Reading Materials	24096.577	2	12048.289	98.181	.00

Within Subjects	235243.73	1917	122.715		
Total	259340.31	1919	135.143		

Table A5 ANOVA Results for Math Scores by Reading Materials (ARTICLES)

Source of Variation	SS	df	MS	F	P
Reading Materials (ARTICLES)	22638.568	2	11319.28	91.227	.00
Within Subjects	238355.10	1921	124.079		
Total	260993.67	1923	1923		

Table A6. ANOVA Results for Math Scores by TELEVISION VIEWING EACH DAY

Source of Variation	SS	df	MS	F	P
TELEVISION VIEWING	20372.210	2	10186.10	81.228	.00
Within Subjects	241146.81	1923	125.401		
Total	261519.02	1925	135.854		

Table A7. ANOVA Results for Math Scores by HOW OFTEN OTHER LANGUAGE SPOKEN IN HOME

Source of Variation	SS	df	MS	F	P
HOW OFTEN OTHER LANGUAGE	3764.657	2	1882.328	14.004	.000
Within Subjects	258215.96	1921	134.417		
Total	261980.61	1923	136.235		

Table A8. ANOVA Results for Math Scores by HOW OFTEN PEOPLE IN HOME SPEAK LANG OTHER

Source of Variation	SS	df	MS	F	P
HOW OFTEN SPEAK LANG OTHER	4017.086	2	2008.543	14.957	.000
Within Subjects	257963.53	1921	134.286		
Total	261980.61	1923	136.235		

Table A10. ANOVA Results for Math Scores by MOTHER'S EDUCATION LEVEL

Source of Variation	SS	df	MS	F	P
MOTHER'S EDUCATION	19197.944	4	4799.486	37.975	.00
Within Subjects	242912.02	1922	126.385		
Total	262109.96	1926	136.090		

Table A11. ANOVA Results for Math Scores by FATHER'S EDUCATION LEVEL

Source of Variation	SS	df	MS	F	P
FATHER'S EDUCATION	26588.600	4	6647.150	54.306	.00
Within Subjects	235009.40	1920	122.401		
Total	261598.00	1924	135.966		

Table A12. ANOVA Results for Math Scores by DOES YOUR FAMILY GET A NEWSPAPER REGULAR

Source of Variation	SS	df	MS	F	P
DOES YOUR FAMILY GET A NEWSPAPER	4639.700	1	4639.700	35.328	.000
Within Subjects	248612.78	1893	131.333		
Total	253252.48	1894	133.713		

Table A13. ANOVA Results for Math Scores by IS THERE AN ENCYCLOPEDIA IN YOUR HOME

Source of Variation	SS	df	MS	F	P
IS THERE AN ENCYCLOPEDIA	2940.423	1	2940.423	22.501	.000
Within Subjects	245545.50	1879	130.679		
Total	248485.93	1880	132.173		

Table A14. ANOVA Results for Math Scores by DOES YOUR FAMILY GET MAGAZINES REGULARLY

Source of Variation	SS	df	MS	F	P
MAGAZINES REGULARLY	11043.835	1	11043.83	87.591	.000
Within Subjects	234768.05	1862	126.084		
Total	245811.88	1863	131.944		

Table A15. ANOVA Results for Math Scores by HOW MUCH TELEVISION DO YOU USUALLY WATCH

Source of Variation	SS	df	MS	F	P
HOW MUCH TELEVISION	22462.365	6	3743.728	30.052	.00
Within Subjects	239056.65	1919	124.574		

Total	261519.02	1925	135.854		

Table A16. ANOVA Results for Math Scores by HOW MUCH TIME EACH DAY IS SPENT ON HOMEWORK

Source of Variation	SS	df	MS	F	P
HOW MUCH TIME ON home work	6563.912	5	1312.782	9.920	.000
Within Subjects	253820.29	1918	132.336		
Total	260384.20	1923	135.405		

Table A17. ANOVA Results for Math Scores by WHAT KIND OF MATH CLASS ARE YOU IN THIS

Source of Variation	SS	df	MS	F	P
KIND OF MATH CLASS	32779.755	2	16389.877	148.504	.00
Within Subjects	197225.09	1787	110.367		
Total	230004.84	1789	128.566		

Table A18. ANOVA Results for Math Scores by HOW MANY DAYS OF SCHOOL MISSED LAST MONTH

Source of Variation	SS	df	MS	F	P
DAYS OF SCHOOL MISSED	3674.025	3	1224.675	9.739	.000
Within Subjects	228495.71	1817	125.754		
Total	232169.74	1820	127.566		

Section 2

Table A19. ANOVA Results for Math Scores by Gender

Source of Variation	SS	df	MS	F	P
Gender	6.707	1	6.707	.072	.788
Within Subjects	183699.021	1974	93.059		
Total	183705.72	1975	93.016		

Table A20: ANOVA Results of Race

Source of Variation	SS	df	MS	F	P
Race	17048.943	2	8524.471	103.775	.00
Within Subjects	153115.28	1864	82.143		

Total	170164.22	1866	91.192		

Table A21. ANOVA Results for Math Scores by SIZE AND TYPE OF COMMUNITY

Source of Variation	SS	df	MS	F	P
SIZE AND TYPE OF COMMUNITY	15063.535	6	2510.589	29.313	.00
Within Subjects	168642.19	1969	85.649		
Total	183705.72	1975	93.016		

Table A22. ANOVA Results for Math Scores by READING MATERIALS

Source of Variation	SS	df	MS	F	P
READING MATERIALS	13189.406	2	6594.703	76.665	.00
Within Subjects	168599.28	1960	86.020		
Total	181788.68	1962	92.655		

Table A23. ANOVA Results for Math Scores by Reading Materials (ARTICLES)

Source of Variation	SS	df	MS	F	P
Reading Materials (ARTICLES)	12490.614	2	6245.307	72.164	.00
Within Subjects	169885.41	1963	86.544		
Total	182376.02	9651	92.812		

Table A24. ANOVA Results for Math Scores by TELEVISION VIEWING EACH DAY

Source of Variation	SS	df	MS	F	P
TELEVISION VIEWING	14339.636	2	7169.818	83.710	.00
Within Subjects	168645.27	1969	85.650		
Total	182984.91	1971	92.839		

Table A25. ANOVA Results for Math Scores by HOW OFTEN OTHER LANGUAGE SPOKEN IN HOME

Source of Variation	SS	df	MS	F	P
OTHER LANGUAGE SPOKEN	2113.825	2	1056.913	11.496	.000
Within Subjects	180844.67	1967	91.939		
Total	182958.50	1969	92.920		

Table A26. ANOVA Results for Math Scores by HOW OFTEN PEOPLE IN HOME SPEAK LANG. OTHER

Source of Variation	SS	df	MS	F	P
HOW OFTEN SPEAK LANG. OTHER	1760.186	2	880.093	9.554	.000
Within Subjects	181198.31	1967	92.119		
Total	182958.50	1969	92.920		

Table A28. ANOVA Results for Math Scores by MOTHER'S EDUCATION LEVEL

Source of Variation	SS	df	MS	F	P
MOTHER'S EDUCATION	15589.952	4	3897.488	45.665	.00
Within Subjects	167713.18	1965	85.350		
Total	183303.13	1969	93.095		

Table A29. ANOVA Results for Math Scores by FATHER'S EDUCATION LEVEL

Source of Variation	SS	df	MS	F	P
FATHER'S EDUCATION	21232.608	4	5308.152	64.330	.00
Within Subjects	161728.10	1960	82.514		
Total	182960.71	182960.71	93.157		

Table A30. ANOVA Results for Math Scores by DOES YOUR FAMILY GET A NEWSPAPER REGULAR

Source of Variation	SS	df	MS	F	P
FAMILY GET A NEWSPAPER	4121.278	1	4121.278	45.996	.000
Within Subjects	173287.63	1934	89.601		
Total	177408.91	1935	91.684		

Table A31. ANOVA Results for Math Scores by IS THERE AN ENCYCLOPEDIA IN YOUR HOME

Source of Variation	SS	df	MS	F	P
ENCYCLOPEDIA IN YOUR HOME	1049.788	1	1049.788	11.439	.001
Within Subjects	176565.17	1924	91.770		
Total	177614.96	1925	92.268		

Table A32. ANOVA Results for Math Scores by DOES YOUR FAMILY GET MAGAZINES REGULARLY

Source of Variation	SS	df	MS	F	P
MAGAZINES REGULARLY	6853.085	1	6853.085	76.748	.000
Within Subjects	170282.80	1907	89.294		
Total	177135.88	1908	92.839		

Table A33. ANOVA Results for Math Scores by HOW MUCH TELEVISION DO YOU USUALLY WATCH

Source of Variation	SS	df	MS	F	P
TELEVISION WATCH	14781.781	6	2463.630	28.781	.00
Within Subjects	168203.12	1965	85.600		
Total	182984.91	1971	92.839		

Table A34 ANOVA Results for Math Scores by HOW MUCH TIME EACH DAY IS SPENT ON HOMEWORK

Source of Variation	SS	df	MS	F	P
TIME EACH DAY SPENT ON HOMEWORK	5206.716	5	1041.343	11.513	.000
Within Subjects	177646.62	1964	90.451		
Total	182853.33	1969	92.866		

Table A35. ANOVA Results for Math Scores by WHAT KIND OF MATH CLASS ARE YOU IN THIS

Source of Variation	SS	df	MS	F	P
KIND OF MATH CLASS	40198.957	2	20099.47	300.120	.00
Within Subjects	121888.02	1820	66.971		
Total	162086.98	1822	88.961		

Table A36. ANOVA Results for Math Scores by HOW MANY DAYS OF SCHOOL MISSED LAST MONTH

Source of Variation	SS	df	MS	F	P
DAYS OF SCHOOL MISSED	4721.446	3	1573.815	18.045	
Within Subjects	162568.94	1864	87.215		

Total	167290.38	1867	89.604		

Section 3

Table A37. ANOVA Results for Math Scores by Gender

Source of Variation	SS	df	MS	F	P
Gender	1401.012	1	1401.012	7.096	.008
Within Subjects	395464.36	2003	197.436		
Total	396865.378	2004	198.037		

Table A38 ANOVA Results for Math Scores by RACE

Source of Variation	SS	df	MS	F	P
RACE	44365.723	2	22182.86	129.322	.00
Within Subjects	322823.18	1882	171.532		
Total	367188.90	1884	194.899		

Table A39. ANOVA Results for Math Scores by SIZE AND TYPE OF COMMUNITY

Source of Variation	SS	df	MS	F	P
SIZE AND TYPE OF COMMUNITY	21637.579	6	3606.263	19.203	.00
Within Subjects	375227.79	1998	187.802		
Total	396865.37	2004	198.037		

Table A40. ANOVA Results for Math Scores by READING MATERIALS

Source of Variation	SS	df	MS	F	P
READING MATERIALS	25598.506	2	12799.25	70.976	.00
Within Subjects	359223.91	1992	180.333		
Total	384822.41	1994	192.990		

Table A41. ANOVA Results for Math Scores by HOME ENVIRONMENT (ARTICLES)

Source of Variation	SS	df	MS	F	P
HOME ENVIRONMENT (ARTICLES)	25405.759	2	12702.87	70.122	.00
Within Subjects	361218.47	1994	181.153		
Total	386624.23	1996	193.700		

Table A42. ANOVA Results for Math Scores by TELEVISION VIEWING EACH DAY

Source of Variation	SS	df	MS	F	P
TELEVISION VIEWING	27026.145	2	13513.07	74.821	.00
Within Subjects	360127.08	1994	180.605		
Total	387153.22	1996	193.965		

Table A43. ANOVA Results for Math Scores by HOW OFTEN OTHER LANGUAGE SPOKEN IN HOME

Source of Variation	SS	df	MS	F	P
OTHER LANGUAGE SPOKEN	7991.279	2	3995.639	20.936	.000
Within Subjects	380560.65	1994	190.853		
Total	388551.93	1996	194.665		

Table A44. ANOVA Results for Math Scores by HOW OFTEN PEOPLE IN HOME SPEAK LANG. OTHER

Source of Variation	SS	df	MS	F	P
SPEAK LANG. OTHER	10399.969	2	5199.984	27.420	.00
Within Subjects	378151.96	1994	189.645		
Total	388551.93	1996	194.665		

Table A46. ANOVA Results for Math Scores by MOTHER'S EDUCATION LEVEL

Source of Variation	SS	df	MS	F	P
MOTHER'S EDUCATION	38124.227	4	9531.057	54.205	.00
Within Subjects	350081.69	1991	175.832		
Total	388205.92	1995	194.589		

Table A47. ANOVA Results for Math Scores by FATHER'S EDUCATION LEVEL

Source of Variation	SS	df	MS	F	P
FATHER'S EDUCATION	48159.983	4	12039.99	70.499	.00
Within Subjects	340025.58	1991	170.781		
Total	388185.56	1995	194.579		

Table A48. ANOVA Results for Math Scores by DOES YOUR FAMILY GET A NEWSPAPER REGULAR

Source of Variation	SS	df	MS	F	P
NEWSPAPER REGULAR	9095.964	1	9095.964	48.294	.000
Within Subjects	369531.74	1962	188.344		
Total	378627.70	1963	192.882		

Table A49. ANOVA Results for Math Scores by IS THERE AN ENCYCLOPEDIA IN YOUR HOME

Source of Variation	SS	df	MS	F	P
ENCYCLOPEDIA IN YOUR HOME	628.234	1	628.234	3.316	.069
Within Subjects	367877.82	1942	189.432		
Total	368506.05	1943	189.658		

Table A50. ANOVA Results for Math Scores by DOES YOUR FAMILY GET MAGAZINES REGULARLY

Source of Variation	SS	df	MS	F	P
Family Gets Magazine Reg.	14045.278	1	14045.27	79.316	.000
Within Subjects	340877.19	1925	177.079		
Total	354922.47	1926	184.280		

Table A51. ANOVA Results for Math Scores by HOW MUCH TELEVISION DO YOU USUALLY WATCH

Source of Variation	SS	df	MS	F	P
HOW MUCH TELEVISION	32298.408	6	5383.068	30.188	.00
Within Subjects	354854.82	1990	178.319		
Total	387153.22	1996	193.965		

Table A52. ANOVA Results for Math Scores by HOW MUCH TIME EACH DAY IS SPENT ON HOMEWORK

Source of Variation	SS	df	MS	F	P
TIME SPENT ON HOMEWORK	7123.673	5	1424.735	7.586	.000
Within Subjects	372224.84	1982	187.803		

Total	379348.51	1987	190.915		

Table A53. ANOVA Results for Math Scores by WHAT KIND OF MATH CLASS ARE YOU IN THIS

Source of Variation	SS	df	MS	F	P
KIND OF MATH CLASS	60907.319	2	30453.65	205.976	.00
Within Subjects	273079.44	1847	147.850		
Total	333986.76	1849	180.631		

Table A54. ANOVA Results for Math Scores by HOW MANY DAYS OF SCHOOL MISSED LAST MONTH

Source of Variation	SS	df	MS	F	P
DAYS OF SCHOOL MISSED	4409.631	3	1469.877	8.214	.000
Within Subjects	332837.67	1860	178.945		
Total	337247.30	1863	181.024		

APPENDIX B

DIF ANALYSES

Table D1. DIF Analyses For Students Who Always Speak a Language Other than English at home (Long-Term Trend Math Test Items, Set 1)

Item	P(Ref)	P(Foc) Diff	Delta Squared	MH Chi-
1	0.97	0.93	-1.81	2.82
2	0.94	0.93	0.64	0.29
3	0.98	0.95	-1.13	0.72
4	0.96	0.96	0.98	0.50
5	0.80	0.83	1.09	3.43
6	0.95	0.94	0.53	0.11
7	0.83	0.76	-0.34	0.23
8	0.88	0.83	-0.18	0.03
9	0.84	0.76	-0.41	0.42
10	0.79	0.71	0.01	0.01
11	0.91	0.87	-0.29	0.08
12	0.76	0.60	-1.02	4.04*
13	0.91	0.83	-1.07	1.94
14	0.74	0.68	0.23	0.13
15	0.64	0.62	0.72	2.13
16	0.72	0.63	-0.26	0.20
17	0.83	0.67	-1.64	9.29**
18	0.60	0.38	-1.41	6.82**
19	0.79	0.69	-0.49	0.86
20	0.30	0.33	0.48	0.88
21	0.69	0.68	1.46	6.44*
22	0.95	0.95	1.05	0.80
23	0.58	0.43	-0.41	0.46
24	0.52	0.40	-0.45	0.76
25	0.63	0.47	-0.74	2.08
26	0.45	0.33	-0.43	0.64
27	0.62	0.70	1.56	9.93**
28	0.95	0.94	0.09	0.01
29	0.59	0.47	-0.14	0.04
30	0.57	0.52	0.75	1.82
31	0.48	0.42	0.64	1.38
32	0.37	0.32	0.36	0.37
33	0.52	0.47	0.26	0.21

34	0.30	0.36	1.84	13.62**
35	0.28	0.22	-0.13	0.02
36	0.29	0.27	0.87	2.15
37	0.26	0.19	0.18	0.03

Table D2. DIF Analyses For Students Who Sometimes Speak a Language Other Than English at Home (Long-Term Trend Math Test Items, Set 1)

Item	P(Ref)	P(Foc) Diff	Delta Squared	MH Chi-
1	0.97	0.96	-0.83	0.74
2	0.94	0.96	0.58	0.53
3	0.98	0.97	-0.74	0.57
4	0.96	0.95	-0.54	0.50
5	0.80	0.81	0.04	0.00
6	0.95	0.94	-0.38	0.26
7	0.83	0.82	-0.46	1.35
8	0.88	0.89	0.31	0.35
9	0.84	0.86	0.09	0.02
10	0.79	0.82	0.27	0.41
11	0.91	0.92	0.15	0.05
12	0.76	0.79	0.06	0.01
13	0.91	0.93	0.94	2.15
14	0.74	0.75	-0.33	0.83
15	0.64	0.66	-0.05	0.01
16	0.72	0.73	-0.16	0.19
17	0.83	0.83	-0.39	0.97
18	0.60	0.61	-0.38	1.19
19	0.79	0.79	-0.14	0.13
20	0.30	0.31	0.01	0.00
21	0.69	0.70	-0.50	1.88
22	0.95	0.95	0.10	0.00
23	0.58	0.69	1.10	10.07**
24	0.52	0.56	0.16	0.24
25	0.63	0.65	-0.01	0.00
26	0.45	0.42	-0.70	5.48*
27	0.62	0.61	-0.27	0.75
28	0.95	0.96	0.18	0.02
29	0.59	0.65	0.42	1.53
30	0.57	0.59	0.00	0.00
31	0.48	0.53	0.23	0.46
32	0.37	0.41	0.22	0.55
33	0.52	0.56	0.22	0.51
34	0.30	0.35	0.36	1.11
35	0.28	0.33	0.31	0.93
36	0.29	0.32	0.00	0.00

37	0.26	0.26	-0.39	1.10

Table D3. DIF Analyses For Students Who Always Speak a Language Other Than English at Home (Long-Term Trend Math Test Items, Set 2)

Item	P(Ref)	P(Foc) Diff	Delta Squared	MH Chi-
1	0.98	0.98	-0.50	0.01
2	0.98	0.98	0.53	0.00
3	0.92	0.92	0.45	0.19
4	0.86	0.80	-0.57	1.01
5	0.61	0.60	0.99	4.14*
6	0.81	0.72	-0.66	1.32
7	0.63	0.60	0.49	0.91
8	0.56	0.53	0.46	0.86
9	0.43	0.45	1.11	5.06*
10	0.49	0.40	-0.37	0.40
11	0.57	0.52	0.45	0.77
12	0.47	0.41	-0.03	0.00
13	0.33	0.37	1.21	5.79*
14	0.41	0.44	1.15	5.67*
15	0.31	0.31	0.41	0.61
16	0.32	0.30	0.24	0.13
17	0.85	0.78	-0.70	1.60
18	0.73	0.62	-0.79	2.54
19	0.79	0.74	0.04	0.00
20	0.74	0.63	-0.73	2.54
21	0.71	0.56	-1.14	6.65**
22	0.51	0.47	0.40	0.54
23	0.43	0.35	-0.52	1.11
24	0.50	0.47	0.42	0.71

Table D4. DIF Analyses For Students Who Sometimes Speak a Language Other Than English at Home (Long-Term Trend Math Test Items, Set 2)

Item	P(Ref)	P(Foc) Diff	Delta Squared	MH Chi-
1	0.98	0.98	-0.79	0.28
2	0.98	0.98	0.06	0.02
3	0.92	0.92	-0.68	1.70
4	0.86	0.87	-0.23	0.25
5	0.61	0.65	-0.07	0.03
6	0.81	0.83	-0.40	0.88
7	0.63	0.68	0.08	0.04
8	0.56	0.59	-0.20	0.37
9	0.43	0.43	-0.44	2.19
10	0.49	0.54	-0.04	0.01
11	0.57	0.61	-0.17	0.23
12	0.47	0.55	0.37	1.33
13	0.33	0.40	0.35	1.12
14	0.41	0.43	-0.27	0.62
15	0.31	0.34	-0.10	0.07
16	0.32	0.36	0.21	0.42
17	0.85	0.85	-0.25	0.37
18	0.73	0.77	0.05	0.01
19	0.79	0.83	0.24	0.35
20	0.74	0.77	0.06	0.01
21	0.71	0.75	0.13	0.13
22	0.51	0.58	0.35	1.23
23	0.43	0.50	0.43	2.34
24	0.50	0.54	-0.11	0.10

Table D5. DIF Analyses For Students Who Always Speak a Language Other Than English at Home (Long-Term Trend Math Test Items, Set 3)

Item	P(Ref)	P(Foc) Diff	Delta Squared	MH Chi-
1	0.96	0.93	-0.53	0.23
2	0.96	0.92	-0.64	0.45
3	0.96	0.96	1.68	1.54
4	0.93	0.92	1.31	2.11
5	0.95	0.88	-1.25	2.49
6	0.90	0.82	-0.66	1.19
7	0.96	0.92	-1.58	3.69
8	0.86	0.69	-1.54	9.64**
9	0.93	0.88	-0.53	0.44
10	0.82	0.76	0.34	0.36
11	0.84	0.78	-0.33	0.39
12	0.86	0.79	-0.51	0.79
13	0.81	0.83	1.14	4.98*
14	0.11	0.09	0.19	0.02
15	0.72	0.69	0.19	0.16
16	0.60	0.56	0.89	3.66
17	0.64	0.56	0.39	0.60
18	0.35	0.29	0.21	0.11
19	0.70	0.56	-0.70	2.40
20	0.62	0.45	-1.06	6.89**
21	0.54	0.41	-0.74	2.90
22	0.38	0.26	-0.83	3.02
23	0.57	0.54	0.97	4.57*
24	0.47	0.38	-0.15	0.07
25	0.56	0.50	0.53	1.34
26	0.48	0.38	-0.04	0.00
27	0.77	0.71	0.02	0.00
28	0.62	0.54	0.17	0.10
29	0.36	0.38	0.63	2.51
30	0.41	0.33	-0.20	0.17
31	0.33	0.37	0.90	4.86*
32	0.24	0.24	0.55	1.31
33	0.46	0.37	-0.34	0.53
34	0.31	0.27	-0.05	0.00
35	0.35	0.30	0.10	0.02
36	0.24	0.24	0.89	3.01

37	0.12	0.14	1.26	4.30*

Table D6. DIF Analyses For Students Who Sometimes Speak a Language Other Than English at Home (Long-Term Trend Math Test Items, Set 3)

Item	P(Ref)	P(Foc) Diff	Delta Squared	MH Chi-
1	0.96	0.96	-0.66	0.67
2	0.96	0.96	0.15	0.00
3	0.96	0.96	-0.09	0.00
4	0.93	0.94	0.21	0.05
5	0.95	0.96	0.22	0.02
6	0.90	0.90	-0.24	0.18
7	0.96	0.96	-0.63	0.68
8	0.86	0.87	0.06	0.00
9	0.93	0.92	-0.69	1.46
10	0.82	0.83	-0.13	0.07
11	0.84	0.84	-0.34	0.82
12	0.86	0.88	0.26	0.30
13	0.81	0.83	0.23	0.34
14	0.11	0.13	-0.10	0.02
15	0.72	0.73	-0.11	0.09
16	0.60	0.62	-0.24	0.56
17	0.64	0.67	0.05	0.01
18	0.35	0.36	-0.52	2.31
19	0.70	0.68	-0.56	2.99
20	0.62	0.67	0.32	1.07
21	0.54	0.55	-0.24	0.61
22	0.38	0.40	-0.15	0.21
23	0.57	0.62	0.23	0.49
24	0.47	0.49	-0.29	0.67
25	0.56	0.62	0.47	2.07
26	0.48	0.52	0.08	0.04
27	0.77	0.82	0.67	3.35
28	0.62	0.70	0.75	5.86*
29	0.36	0.37	-0.05	0.02
30	0.41	0.44	0.07	0.04
31	0.33	0.33	-0.28	0.86
32	0.24	0.27	0.12	0.11
33	0.46	0.52	0.34	1.38
34	0.31	0.36	0.15	0.18
35	0.35	0.39	-0.01	0.00
36	0.24	0.31	0.46	1.75

37	0.12	0.15	0.18	0.13

Table D7. DIF Analyses For Students Who Always Speak a Language Other Than English at Home (Long-Term Trend Science Test Items, Set 1)

Item	P(Ref)	P(Foc) Diff	Delta Squared	MH Chi-
1	0.64	0.57	0.18	0.07
2	0.88	0.81	-0.12	0.00
3	0.63	0.58	0.41	0.67
4	0.80	0.71	-0.41	0.52
5	0.74	0.58	-0.76	2.25
6	0.39	0.35	0.05	0.00
7	0.65	0.57	-0.22	0.15
8	0.94	0.81	-1.73	6.11*
9	0.75	0.58	-0.80	2.11
10	0.54	0.40	-0.50	0.90
11	0.56	0.47	0.14	0.04
12	0.43	0.32	-0.31	0.28
13	0.55	0.54	0.70	2.19
14	0.40	0.35	0.36	0.42
15	0.36	0.32	0.72	1.66
16	0.62	0.48	-0.38	0.48
17	0.59	0.58	1.06	4.32*
18	0.42	0.45	0.95	4.25*
19	0.47	0.36	0.03	0.00
20	0.37	0.32	0.46	0.77
21	0.26	0.18	-0.35	0.25
22	0.15	0.15	0.43	0.35
23	0.18	0.17	0.52	0.56
24	0.14	0.09	-0.4	0.15

Table D8. DIF Analyses For Students Who Sometimes Speak a Language Other Than English at Home (Long-Term Trend Science Test Items, Set 1)

Item	P(Ref)	P(Foc) Diff	Delta Squared	MH Chi-
1	0.64	0.61	-0.55	3.59
2	0.88	0.91	0.47	0.85
3	0.63	0.65	0.00	0.00
4	0.80	0.81	-0.18	0.22
5	0.74	0.74	-0.40	1.37
6	0.39	0.42	0.25	0.84
7	0.65	0.66	-0.02	0.00
8	0.94	0.92	-0.73	1.24
9	0.75	0.79	0.34	0.72
10	0.54	0.60	0.34	1.11
11	0.56	0.57	-0.15	0.22
12	0.43	0.48	0.37	1.55
13	0.55	0.58	0.17	0.31
14	0.40	0.41	-0.12	0.14
15	0.36	0.39	0.03	0.00
16	0.62	0.65	0.09	0.05
17	0.59	0.64	0.28	0.75
18	0.42	0.48	0.60	4.95*
19	0.47	0.52	0.14	0.16
20	0.37	0.38	-0.15	0.23
21	0.26	0.24	-0.41	1.48
22	0.15	0.18	0.43	1.37
23	0.18	0.17	-0.48	1.63
24	0.14	0.14	-0.14	0.08

Table D9. DIF Analyses For Students Who Always Speak a Language Other Than English at Home (Long-Term Trend Science Test Items, Set 2)

Item	P(Ref)	P(Foc) Diff	Delta Squared	MH Chi-
1	0.65	0.56	-0.23	0.22
2	0.73	0.67	0.04	0.00
3	0.69	0.62	0.35	0.50
4	0.45	0.36	-0.07	0.01
5	0.67	0.56	-0.24	0.24
6	0.61	0.47	0.36	0.45
7	0.87	0.76	-0.64	1.09
8	0.74	0.58	-0.37	0.45
9	0.69	0.57	0.48	0.78
10	0.46	0.34	-0.15	0.06
11	0.72	0.69	0.59	1.43
12	0.62	0.50	-0.26	0.30
13	0.90	0.87	0.61	0.64
14	0.63	0.62	0.40	0.72
15	0.79	0.66	-0.44	0.68
16	0.83	0.78	0.26	0.15
17	0.62	0.51	-0.37	0.69
18	0.63	0.54	-0.11	0.03
19	0.61	0.58	0.53	1.47
20	0.49	0.53	1.08	6.48*
21	0.20	0.16	0.59	0.83
22	0.54	0.48	0.16	0.09
23	0.52	0.42	-0.43	0.89
24	0.43	0.44	0.87	4.19*
25	0.38	0.25	-0.54	1.06
26	0.37	0.32	0.25	0.24
27	0.46	0.44	0.16	0.09
28	0.34	0.20	-0.87	2.38
29	0.32	0.27	0.11	0.02
30	0.45	0.38	0.09	0.02
31	0.19	0.14	-0.10	0.00

Table D10. DIF Analyses For Students Who Sometimes Speak a Language Other Than English at Home (Long-Term Trend Science Test Items, Set 2)

Item	P(Ref)	P(Foc) Diff	Delta Squared	MH Chi-
1	0.65	0.70	0.23	0.58
2	0.73	0.77	0.23	0.43
3	0.69	0.72	0.07	0.03
4	0.45	0.52	0.34	1.51
5	0.67	0.70	-0.04	0.01
6	0.61	0.62	-0.54	2.62
7	0.87	0.90	0.10	0.02
8	0.74	0.76	-0.25	0.42
9	0.69	0.73	-0.05	0.00
10	0.46	0.51	0.07	0.03
11	0.72	0.71	-0.53	3.14
12	0.62	0.63	-0.24	0.64
13	0.90	0.93	0.64	1.33
14	0.63	0.65	-0.06	0.02
15	0.79	0.74	-1.16	12.31**
16	0.83	0.86	0.15	0.09
17	0.62	0.65	0.05	0.01
18	0.63	0.66	0.10	0.09
19	0.61	0.64	0.03	0.00
20	0.49	0.50	-0.18	0.38
21	0.20	0.22	0.01	0.00
22	0.54	0.60	0.25	0.75
23	0.52	0.52	-0.27	1.00
24	0.43	0.48	0.28	1.03
25	0.38	0.44	0.31	1.02
26	0.37	0.37	-0.37	1.56
27	0.46	0.42	-0.61	5.36*
28	0.34	0.39	0.21	0.46
29	0.32	0.31	-0.44	2.28
30	0.45	0.50	0.11	0.14
31	0.19	0.22	0.12	0.10

Table D11. DIF Analyses For Students Who Always Speak a Language Other Than English at Home (Long-Term Trend Science Test Items, Set 3)

Item	P(Ref)	P(Foc) Diff	Delta Squared	MH Chi-
1	0.50	0.34	-0.76	3.02
2	0.86	0.76	-0.61	1.27
3	0.89	0.78	-1.06	3.90*
4	0.86	0.72	-1.02	3.23
5	0.71	0.65	-0.02	0.00
6	0.76	0.67	-0.18	0.10
7	0.44	0.44	0.72	3.11
8	0.72	0.74	0.80	3.12
9	0.61	0.62	0.47	1.32
10	0.56	0.51	0.13	0.06
11	0.66	0.66	0.87	4.09*
12	0.93	0.85	-0.86	1.40
13	0.58	0.52	0.41	0.91
14	0.58	0.54	0.29	0.44
15	0.66	0.54	-0.54	1.49
16	0.85	0.71	-1.13	6.07*
17	0.62	0.54	0.23	0.22
18	0.44	0.48	0.96	6.19*
19	0.53	0.44	-0.27	0.37
20	0.60	0.53	0.24	0.23
21	0.36	0.39	0.23	0.32
22	0.47	0.41	-0.17	0.15
23	0.22	0.22	-0.13	0.05
24	0.24	0.17	-0.45	0.66
25	0.41	0.31	-0.45	1.01
26	0.23	0.16	-0.33	0.29
27	0.53	0.49	0.30	0.52

Table D12. DIF Analyses For Students Who Sometimes Speak a Language Other Than English at Home (Long-Term Trend Science Test Items, Set 3)

Item	P(Ref)	P(Foc) Diff	Delta Squared	MH Chi-
1	0.50	0.48	-0.41	1.99
2	0.86	0.88	0.25	0.25
3	0.89	0.89	-0.19	0.11
4	0.86	0.91	1.08	4.58*
5	0.71	0.71	-0.02	0.00
6	0.76	0.77	-0.13	0.12
7	0.44	0.45	0.00	0.00
8	0.72	0.74	0.04	0.00
9	0.61	0.65	0.18	0.39
10	0.56	0.59	0.10	0.11
11	0.66	0.73	0.69	4.98*
12	0.93	0.94	-0.40	0.34
13	0.58	0.63	0.52	3.14
14	0.58	0.58	-0.12	0.17
15	0.66	0.66	-0.21	0.45
16	0.85	0.83	-0.63	2.60
17	0.62	0.67	0.41	1.88
18	0.44	0.45	0.01	0.00
19	0.53	0.57	0.29	1.03
20	0.60	0.63	0.07	0.03
21	0.36	0.34	-0.19	0.43
22	0.47	0.46	-0.19	0.49
23	0.22	0.22	-0.09	0.05
24	0.24	0.28	0.27	0.65
25	0.41	0.40	-0.31	1.08
26	0.23	0.23	-0.11	0.08
27	0.53	0.61	0.72	6.67**