RESEARCH REPORT 2013-6

CollegeBoard

Differential Item and Person Functioning in Large-Scale Writing Assessments Within the Context of the SAT®

By George Engelhard Jr., Stefanie A. Wind, Jennifer L. Kobrin, and Michael Chajewski

George Engelhard Jr. is a professor of educational measurement and policy in the Department of Educational Psychology at The University of Georgia in Athens, Ga.

Stefanie A. Wind is a doctoral student in educational measurement and policy in the Division of Educational Studies at Emory University in Atlanta, Ga.

Jennifer L. Kobrin was a research scientist at the College Board.

Michael Chajewski is an associate psychometrician at the College Board.

Acknowledgments

An earlier version of this paper was presented at the annual meeting of the American Educational Research Association in Vancouver, April 2012. Rebecca Zwick (University of California, Santa Barbara) provided a critical review and helpful feedback on a draft of our research report. The College Board provided support for this research. The College Board encourages researchers to freely express their professional judgments. Therefore, the points of view or opinions stated in College Board–supported research do not necessarily represent official College Board positions or policies.

About the College Board

The College Board is a mission-driven not-for-profit organization that connects students to college success and opportunity. Founded in 1900, the College Board was created to expand access to higher education. Today, the membership association is made up of over 6,000 of the world's leading educational institutions and is dedicated to promoting excellence and equity in education. Each year, the College Board helps more than seven million students prepare for a successful transition to college through programs and services in college readiness and college success — including the SAT® and the Advanced Placement Program®. The organization also serves the education community through research and advocacy on behalf of students, educators and schools. For further information, visit www.collegeboard.org.

RESEARCH

For more information on College Board research and data, visit research.collegeboard.org.

Contents

Tables

Figures

Executive Summary

The purpose of this study is to illustrate the use of explanatory models based on Rasch measurement theory to detect systematic relationships between student and item characteristics and achievement differences using differential item functioning (DIF), differential group functioning (DGF), and differential person functioning (DPF) techniques. The major focus of the analyses in this study was to demonstrate a set of methodological techniques that can be used to better understand subgroup performance on a large-scale writing assessment, rather than to conduct bias or sensitivity reviews. DIF, DGF, and DPF are conceptualized as types of model-data misfit to a Rasch measurement model. Specifically, the SAT® writing section (SAT-W) is used to illustrate this perspective on DIF, DGF, and DPF. Although the current analyses that are in place to examine reliability, validity, and fairness related to the SAT-W are sufficient for examining the psychometric quality of this assessment, the analyses serve as additional tools that supplement the routine analyses. The substantive research questions examine whether selected student characteristics (gender, race/ethnicity, and best language) influence DIF, and also whether subgroups of students function differentially on different SAT-W item subsets (sentence correction: 25 items; usage: 18 items; and revision in context: six items). Data analyses were conducted with the Facets computer program (Linacre, 2007). A random sample of students from the October 2009 administration of the SAT was used in this study (n = 19,341).

The results of the study suggest that the SAT-W items exhibit very good model-data fit to the Rasch measurement model. As found in previous research on writing, there were small subgroup differences, with females having a higher level of writing achievement than males. The Asian, Asian American, or Pacific Islander subgroup had the highest level of writing achievement, while the black or African American subgroup had the lowest level. In terms of best-language subgroups, the English only subgroup had the highest level of writing achievement. Overall, there did not appear to be any item subsets functioning in an unexpected way across the subgroups of persons (gender, race/ethnicity, and best-language subgroups). The results of the differential person functioning analyses indicate that some individuals did not respond to the SAT-W items as expected based on the Rasch measurement model. A promising area for future research is to examine within-person variation in responding to items on the SAT-W.

Introduction

Writing is an essential aspect of communicative competence in modern societies (Behizadeh and Engelhard, 2011; Elliot, 2005). In the United States, for example, the new Common Core Standards Initiative (2010) stresses an integrated model of literacy:

> Although the Standards are divided into Reading, Writing, Speaking and Listening, and Language strands for conceptual clarity, the processes of communication are closely connected, as reflected throughout this document. For example, Writing Standard 9 requires that students be able to write about what they read. Likewise, Speaking and Listening Standard 4 sets the expectation that students will share findings from their research. (p. 4)

Writing is considered a key ingredient for college and career success. Many universities require essays and other evidence of writing competence for admission to higher education. The SAT is one of the most widely used college admission assessment systems, and a writing section (SAT-W) was added in March 2005 (Kobrin & Kimmel, 2006; Mattern, Camara, & Kobrin, 2007). The purpose of the overall SAT is to assess the critical reading, mathematical reasoning, and writing skills that students have developed over time and that they need to be successful in college. The essay is designed to provide evidence that students can develop a point of view on an issue presented in an excerpt, support their point of view using reasoning and examples from reading, studies, experience, or observations, and follow the conventions of standard written English. The other three sections of the SAT-W consist of objective, or selected-response, items designed to assess student skills in sentence correction, usage, and revision in context based on the conventions of standard written English.

This study focuses on model-data fit as a type of validity evidence for the SAT-W from the perspective of modern item response theory using the many-facet Rasch (MFR) model (Linacre, 2007). As pointed out by Messick (1995),

> Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. These scores are a function not only of the items or stimulus conditions, but also of the persons responding as well as the context of the assessment. In particular, what needs to be valid is the meaning or interpretation of the score; as well as any implications for action that this meaning entails (Cronbach, 1971). The extent to which score meaning and action implications hold across persons or population groups and across settings or contexts is a persistent and perennial empirical question. (p. 741)

Current work on the concept of validity stresses the use of test scores (Kane, 1992, 2001), and the development of evidence-centered designs

This study focuses on model-data fit as a type of validity evidence for the SAT-W from the perspective of modern item response theory using the many-facet Rasch (MFR) model (Linacre, 2007).

to support validity arguments (Huff, Steinberg, & Matts, 2010; Mislevy, Steinberg, Breyer, Almond, & Johnson 2002). These aspects of validity tell only part of the story. As pointed out by Messick (1995), validity studies should also address "score meaning," and explicitly recognize that score meaning is a function of persons and items, as well as of contextual aspects of the assessment. Modern item response theory supports an evaluation of score inferences that explicitly recognizes that score meaning is a function of items, persons, and context (Embretson, 1996). In particular, invariant measurement (Engelhard, 2009) provides a coherent approach to differential item functioning and differential person functioning within the context of persons, person subgroups, items, and item subsets.

Previous research has been conducted on differential item functioning (DIF) related to the SAT within the mathematics, verbal/critical reading, and writing sections (Curley & Schmitt, 1993). There has been less published research on DIF within the context of the SAT-W in comparison to other sections of the SAT, although routine DIF analyses are a part of the standard test development process for the SAT-W that are sufficient for examining the psychometric quality of these procedures. This study uses Rasch Measurement Theory to explore model-data fit on the SAT-W from the perspective of invariant measurement (Engelhard, 2009). Specifically, model-data fit and residual analyses are conducted using item response functions (differential item functioning), group response functions (differential group functioning), and person response functions (differential person functioning).

The next section describes the concept of invariant measurement, followed by a section that summarizes relevant research on subgroup differences in writing.

Invariant Measurement

Invariance is a fundamental concept in measurement (Engelhard, 2013; Engelhard and Perkins, 2011; Millsap, 2011). The goal of developing instruments that facilitate invariant measurement has deep historical roots in the human sciences (Engelhard, 2008). An assessment's capacity to provide invariant measures is not directly observable, but evidence can be evaluated based on item and person fit indexes that provide warrants for the claims of invariant measurement (Engelhard, 2009). Wright (1968) described the requirements for invariant measurement based on Rasch's idea of specific objectivity. Engelhard (2009) extended these requirements, and proposed a framework for examining differential item functioning and differential person functioning based on the concept of invariant measurement. The basic requirements of invariant measurement can be summarized in terms of item calibrations, person measurements, and a variable map:

Item calibration:

1. Person-invariant calibration of test items: The calibration of the items must be independent of the particular persons used for calibration.

2. Noncrossing item response functions: Any person must have a better chance of success on an easy item than on a more difficult item.

Person measurement:

3. Item-invariant measurement of persons: The measurement of persons must be independent of the particular items that happen to be used for the measuring.

4. Noncrossing person response functions: A more able person must always have a better chance of success on an item than a less able person.

Variable map:

5. Unidimensionality: Person and items must be located on a single underlying latent variable.

Requirements 1 and 2 are related to DIF, while requirements 3 and 4 address issues related to DPF. The fifth requirement is fundamental for creating a visual display that illustrates the construct represented by the assessment. As discussed further below, adherence to the fifth requirement highlights the practical utility of invariant measurement: When a construct can be illustrated as a single line, item difficulties can be interpreted independently from a particular sample of persons, and person achievement measures can be interpreted independently from a particular sample of items.

These requirements lay the foundation for conceptualizing DIF in terms of a failure to meet the requirements of person-invariant item calibration (items do not have the same location on the latent variable for different persons). It also suggests the view that DPF is related to item-invariant measurement of persons (persons do not have the same interpretations of different items). It is well known that invariant measurement is only obtained when there is good model-data fit for both items and persons (Swaminathan, Hambleton, & Rogers, 2007). According to Hambleton (1989),

> The potential of item response theory for solving many problems in testing and measurement is high; however, the success of particular IRT applications is not assured simply by processing test results through one of the available computer programs. The advantages claimed for item response models can be realized only when the fit between the model and the test data set of interest is satisfactory. A poorly fitting model cannot yield invariant item- and ability-parameter estimates. (p. 172)

Research on DIF can be viewed as an examination of the claim of person-invariant calibration of items, or measurement invariance. In a parallel fashion, research on DPF can be defined as the identification of unexpected differences between observed and model-expected performance of persons on a set of items that influence the meaning of a test score. Studies of DPF can be viewed as exploratory models testing the hypothesis of item-invariant measurement of persons. In addition to examining model-data fit at the level of items and persons, DGF can also be used as exploratory models for examining person subgroups (gender, race/ethnicity, and best language) and item subsets.

A variety of methods are employed to examine differential performance by groups of students on tests or individual test items. Differences in achievement have been attributed to actual differences unrelated to the particular characteristics of a test or individual items, differences related to an external (nontest) variable that affects test performance, or both. Zumbo (2007) differentiates three concepts related to the analysis of group performance differences on test items: item impact, differential item functioning (DIF), and item bias. In looking at subgroups, it is important to keep in mind the distinctions among impact, DIF, and bias. Zumbo (2007) defines these terms as follows:

- Item impact: Item impact is evident when examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item because there are true differences between the groups in the underlying ability being measured by the item.
- Differential item functioning: DIF occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item *after matching on the underlying ability* that the item is intended to measure.

Because the concepts of gender, race, and best language are not easily defined, identification and analysis of probable sources for the differential performance between these subgroups of test-takers is complex.

• Item bias: Item bias occurs when examinees of one group are less likely to answer an item correctly (or endorse an item) than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose. DIF is required, but not sufficient, for item bias (p. 12, italics in original).

With regard to writing assessment, research has examined achievement gaps in order to evaluate and address issues of fairness in assessment and instructional environments, as well as in the development of new standards and curricula that guide both environments (Noeth & Kobrin, 2007). Research on large-scale writing assessments includes analyses of the content, format, and administration procedures of assessments whose results indicate disparate performance between groups of students. Because the concepts of gender, race, and best language are not easily defined, identification and analysis of probable sources for the differential performance between these subgroups of test-takers is complex. In general, this research reflects descriptive, post-hoc analyses of differential performance across student subgroups.

Writing Achievement and Student Subgroups

The next three sections summarize research on writing achievement related to student gender, race/ethnicity, and student self-reports of their best language.

Writing Achievement and Student Gender

Gaps in writing achievement by gender have been examined in terms of the various prompts and writing tasks used in large-scale assessments. Research on the impact of SAT-W prompt types (Breland, Kubota, Nickerson, Trapani, & Walker, 2004) and placement of prompts (Oh & Walker, 2006) on scores indicates that female students tend to receive higher scores than male students regardless of prompt type and placement. Breland et al. (2004) found statistically significant gender differences across all prompts ($p < 0.05$), with effect sizes ranging from $d = 0.19$ to $d = 0.31$. Findings by Oh and Walker (2006) also indicated significant differences ($p < 0.001$) between female and male achievement on SAT-W essays regardless of prompt placement and type. In general, this research suggests that female test-takers can be expected to score higher than male test-takers on the SAT-W.

Along the same lines, Engelhard, Gordon, and Gabrielson (1992) found gender to be a significant predictor of writing achievement across a variety of writing tasks on a statewide writing assessment. In their sample of eighth-grade students, females performed significantly higher than males regardless of the mode of discourse or experiential demand required by a

writing task. Research indicates that females outperform males on writing assessments (e.g., Cole, 1997, 2000; Mattern, Camara, & Kobrin, 2007), and that additional research that focuses on DIF and DPF, such as this study, may illuminate other aspects of gender differences in writing achievement.

Persistent differences in gender achievement on large-scale writing tests have also been examined at analytic or domain levels. These studies tend to reveal similar patterns of male and female performance to research on holistic writing scores. When compositions are examined at the domain level, females have been found to outperform male students in score categories related to both meaning (e.g., style and organization), and mechanics (e.g., conventions and sentence formation) of writing (Engelhard et al., 1992). Breland, Bonner, and Kubota (1995) examined correlations among a variety of analytically scored features with overall scores from the 1990 administration of the English Composition Test. The English Composition Test is a mixed-format assessment of writing competence that was once part of the SAT Subject Tests™. Analytic scoring reveals similar patterns for male and female students, with females generally outperforming males across domains.

Engelhard, Gordon, and Gabrielson (1992) found a stronger gender effect within mechanics and usage domains (effect sizes were $d = 0.49$, and $d = 0.39$, respectively), than within content/organization, style, and sentence formation domains ($d = 0.33$, $d = 0.33$, and $d = 0.36$, respectively). Engelhard, Gordon, Walker, and Gabrielson (1994) obtained similar findings in an examination of nearly 171,000 Georgia eighth-grade students. As in the 1992 study, the gender effect indicating higher scores for female students was greater for conventions domains than content domains.

Writing Achievement and Student Race/Ethnicity

Along with gender, achievement gaps related to race/ethnicity are a topic of intense interest in assessment research. Although research tends to show higher achievement trends for "majority" than "minority" groups (Engelhard et al., 1994; Breland et al., 2004), patterns of subgroup differences have also been shown to vary across subject areas and for different types of writing prompts (Pomplun, Wright, Oleka, & Sudlow, 1992; Breland et al., 1999). In general, research on the SAT-W indicates higher performance by white and Asian students than by other racial/ethnic subgroups (Mattern, Camara, & Kobrin, 2007; Sathy, Barbuti, & Mattern, 2006).

Several studies have investigated essay-based assessments in terms of the interaction between prompt types and student characteristics, including race/ethnicity. In a preliminary

Research indicates that females outperform males on writing assessments (e.g., Cole, 1997, 2000; Mattern, Camara, & Kobrin, 2007), and that additional research that focuses on DIF and DPF, such as this study, may illuminate other aspects of gender differences in writing achievement.

research study for the essay section of the SAT, Breland et al. (2004) investigated differences in racial/ethnic subgroup performance across four prompts in order to identify interactions between performances and prompt characteristics. Two traditional SAT Subject Test in Writing prompts were used, along with two modified prompts that encouraged persuasive writing. Findings indicated significant score gaps across racial/ethnic subgroups, and Breland et al. (2004) found that differences were persistent across prompt types, and thus could not be attributed directly to prompt characteristics. Overall, trends in this study matched those of other studies of essay writing assessments, and potential causes for these subgroup performance differences are not easily identified.

Differences in racial/ethnic subgroup performance on large-scale writing tests have also been considered as they appear across sections of analytic essay rubrics. In their analysis of the 1990 administration of the English Composition Test that was mentioned earlier, Breland, Bonner, and Kubota (1995) examined essay scores at both the holistic- and analytic-score level for racial/ethnic subgroups. At the analytic level, essay features related to organization were most strongly correlated with high holistic scores across subgroups, but differences were found in each of the second-strongest correlates for Asian American, black, Hispanic, and white students. Along the same lines, Engelhard, Gordon, Walker, and Gabrielson (1994) investigated domain-level performance for black and white eighth-grade students in a statewide writing assessment, and examined subgroup differences in terms of domain categories (content/organization, style, sentence formation, usage, and mechanics), response mode (narrative, descriptive, and expository), and experiential demand (direct experience, imagined experience, and outside knowledge). According to the authors: "the fact that observed differences between black students and white students continue to be evident and significant on the mechanics portion of the eighth-grade assessment suggests that many black students may not have been given the appropriate opportunities in school to master the necessary code-switching skills or the ability to transition between language patterns used in and out of school" (p. 207).

Writing Achievement and Student Language

Achievement differences on both multiple-choice and essay-based writing assessments related to language and English proficiency have been a focus in assessment research since around the 1980s (Cumming, 2001). Research on language-related achievement gaps is widespread, with studies examining performance trends related to overall performance (Crane, Barrat, & Huang, 2011), classroom placement, age, and number of years spent in primarily English-speaking countries (Tarone et al., 1993), writing processes (Fitzgerald, 2006; Hamp-Lyons, 1991). and distinct textual features of compositions written by students with multilingual literacy practices (Bermúdez & Prater, 1994; Carlisle & McKenna, 1990; Hinkel, 2003; Silva, 1993; Vann, Lorenz, & Meyer, 1991).

Similar to research on writing achievement by gender and racial/ethnic subgroups, previous studies on the performance of language groups on the SAT-W are mainly descriptive, and focus on achievement trends and differential item and prompt impact. In general, this research has concluded that students whose best language is English perform better on the SAT-W than other language groups (Sathy, Barbuti, & Mattern, 2006). In their analysis of group performance across SAT-W prompts, Breland et al. (2004) noted that language group differences in achievement persist across prompt types. They found statistically significant differences between essay performance by English Best Language (EBL) students and English Not Best Language (ENBL) students on four different prompts ($p < 0.05$), with effect sizes ranging from $d = 0.36$ to $d = 0.66$. Similarly, Oh and Walker (2006) found that EBL students tend to perform higher than ENBL students on the SAT-W regardless of prompt

placement within the test or prompt type. Differences in essay scores between these two groups were statistically significant ($p < 0.001$), and no interaction effects were found between language groups and essay placement or prompt type; these findings suggest a persistent trend of higher performance by EBL students on the essay portion of the SAT-W when compared to ENBL students.

In addition to analyses of overall scores, studies related to the writing achievement of language groups have also examined performance at the analytic, or domain-score, level. A notable study by Tarone et al. (1993) examined writing achievement by eighth-, 10th-, and 12th-grade EBL and ENBL Southeast Asian American students. High correlations were found across domain-level scores for both groups of students, indicating similar or related performance by these subgroups across various parts of an analytic essay rubric. It is important to note that this study found no significant difference in scores assigned to eighth-, 10th-, or 12th-grade ENBL students. These authors attributed the apparent lack of improvement to the fact that ENBL students had different opportunities to receive writing instruction than did EBL students.

Research that seeks to explain differences in the writing achievement of language groups has also examined the nature of ENBL compositions, along with writing processes used by these students. Numerous literature reviews and meta-analytic studies summarize empirical research findings related to these variables. For example, Silva (1993) examined 72 studies on differences in the composing processes, features, and structures of compositions across samples of students who represented 27 best languages besides English. Although this study revealed findings of broad similarity between EBL and ENBL compositions at a holistic level, consideration of compositions in terms of individual features indicated that essays composed by ENBL students are "strategically, rhetorically, and linguistically different in important ways from [EBL] writing" (p. 669). Along the same lines, Hinkel (2003) examined writing samples from EBL and ENBL students in terms of syntactic and lexical simplicity and complexity, and found evidence of a "restricted lexical repertoire" for ENBL students (p. 293). Similarly, Fitzgerald (2006) considered writing competence in terms of writing process development and essay features in a literature review of research on multilingual writing practices of students from preschool to 12th grade. She was unable to draw firm conclusions regarding the nature of differences between writing practices and competence related to best language. A persistent theme across these reviews, in the words of Fitzgerald (2006), is that "second-language writing ability looms large in students' academic development" and "is critical to educational advancement and future opportunities" (pp. 351–352). The persistent achievement differences by language subgroups highlight the need for research that examines interactions between assessment and student characteristics.

Similar to research on writing achievement by gender and racial/ethnic subgroups, previous studies on the performance of language groups on the SAT-W are mainly descriptive, and focus on achievement trends and differential item and prompt impact.

In summary, previous research on subgroup differences in writing achievement suggests that, in general, there are fairly consistent subgroup differences that tend to indicate higher scores for whites, females, and individuals for whom English is their best language. The current study extends this research by drilling down more deeply into interactions among item subsets and subgroup membership with individual and subgroup-level analyses of person fit. The focus on differential item and person functioning in regard to item subsets and person subgroups holds promise to add to the literature in these areas.

Purpose

The purpose of this study is to explore differential item functioning (DIF), differential person functioning (DPF), and differential group functioning (DGF) within the context of large-scale writing assessments. The main goal of this study is to examine the model-data fit of item calibrations and person measurements in terms of student subgroups and item subsets on the SAT-W. Two research questions are used to guide the analyses:

1. Are items functioning differentially for subgroups of persons (i.e., gender, race/ethnicity, and best language)?
2. Are persons and subgroups responding as intended to different subsets of items (i.e., sentence correction, usage, and revision in context)?

The focus on differential item and person functioning in regard to item subsets and person subgroups holds promise to add to the literature in these areas.

Through the examination of these research questions, the analyses in this study illustrate methods for more fully understanding how differential item, person, and group functioning, respectively, are related to writing achievement as measured by the SAT-W.

Method

Instrument

The SAT-W was introduced in 2005. The SAT writing section consists of 49 multiple-choice items classified into three types: (1) sentence correction (25 items), (2) usage (18 items), and (3) revision in context (six items). Students also respond to a 25-minute essay. In this study, the multiple-choice items are scored as 1s if answered correctly ($x = 1$), scored as 0s if answered incorrectly or classified as omitted ($x = 0$), and coded as "missing" if the student did not reach the item ($x = .$).[1] Two raters score each essay in seven categories (0 to 6) using the rubric shown in Appendix A. A score of 0 is reserved for students who do not write an essay, essays written on a topic that was not addressed in the prompts, or extremely illegible essays that are not scorable.

1. In calculation of the item score for each student, omitted items (where students did not respond to that item but did respond to at least one item placed later in the test) were scored as incorrect, and items that were not reached (where students did not respond to that item or any item placed later in the test) were designated as missing.

Participants

Table 1 presents the descriptive information for the total population (N = 388,889) and a random sample of 5 percent of the test-takers who were used in this study. The random sample of students is from the October 2009 SAT administration (n = 19,341). As expected, the data in Table 1 support the inference that there is a close match between the demographic characteristics of the total population and the random sample included in this study.

Procedures

Data analyses were conducted with the Facets computer program (Linacre, 2007), and three models based on Rasch measurement theory were used to analyze the data. In this section, each model is described in terms of its relationship to the research questions.

Two-facet partial credit model. The two-facet partial credit (PC) model (Wright & Masters, 1982) is a generalization of the Rasch model for dichotomous data, and it can be applied to rating scale data in two or more ordered categories. In the context of the SAT-W, both dichotomously scored multiple-choice items and essay ratings can be modeled together. The PC model allows for variation in the number of categories used by raters to score the essay items, and can be used to identify differences in rater use of rating scale categories. The PC model can be expressed mathematically as:

$$ln\,[P_{nik} / P_{nik-1}] = \theta_n - \delta_i - \tau_{ik},$$

where

P_{nik} = probability of person n on item i scoring k;
P_{nik-1} = probability of person n on item i scoring k–1;
θ_n = location of person n on latent variable;
δ_i = difficulty of item i; and
τ_{ik} = difficulty of moving from category k–1 to k within item i.

There are several things to note. First, the PC model can be easily modified to examine person subgroups and item subsets by adding various interaction effects to the general model. Second, the thresholds, τ_{ik}, are defined as zero for the dichotomous items and the category coefficients for the rating categories used to score the essays. Finally, once the parameters of the model have been estimated, residual analyses can be used to explore item and person response behaviors in detail.

After estimates of the main-effect parameters are computed, several statistics can be examined to identify further characteristics of the data that are of interest. First, the *reliability of separation* statistic based on Rasch models is an index of how well individual elements within a facet can be differentiated from one another, such as individual persons or items.

The reliability of separation statistics for persons is comparable to Cronbach's coefficient alpha and KR20 because it reflects an estimate of true score to observed score variance. For the other facets, the reliability of separation statistic describes the spread or differences between elements within a facet, such as differences in rater severity. The statistic is calculated as follows:

$$Rel = (SD^2 - MSE) / SD^2,$$

where SD^2 is the observed variance of elements within a facet in logits and MSE is the mean square calibration error. MSE is estimated as the average value of calibration error variances

(squares of the standard errors) for each element within a facet. Andrich (1982) provides a detailed derivation of this reliability of separation index.

Next, residuals can be obtained based on the difference between observed and expected values from the model. Fit statistics are used within Rasch-based approaches in order to examine the degree to which adherence to the requirements for invariant measurement is observed in a set of data. Model-data fit analyses within Rasch measurement theory typically focus on fit statistics that summarize residuals, or differences, between model expectations and empirical observations. In the context writing assessment, fit statistics can be used to identify multiple-choice items and students that do not match the expectations of the ideal-type model. This study focuses on two fit statistics that are computed in the Facets program (Linacre, 2007): *infit* and *outfit* statistics. These statistics can be calculated for facets related to persons and items, as well as for other explanatory facets included in the model.

Outfit is calculated by summing standardized residual variance across facets. Because it is unweighted, the outfit statistic is useful because it is particularly sensitive to outliers, or extreme unexpected observations. The person outfit (U_n) statistic is calculated as follows:

$$U_n = \frac{\sum_{i=1}^{L} Z_{ni}^2}{L} \quad ,$$

where Z^2_{ni} represents standardized score residuals and L is the number of items. Similarly, the outfit statistic for items (U_i) is calculated as:

$$U_i = \frac{\sum_{n=1}^{N} Z_{ni}^2}{N} \quad ,$$

where N is the number of persons.

Infit statistics are also useful for evaluating model-data fit, but are less sensitive to outlying data because residuals are weighted by the variance of an individual facet, which reduces the impact of unexpected observations. Similar to outfit, infit can be calculated for person- and item-related facets. The infit statistic for persons (U_n) is calculated as:

$$U_n = \frac{\sum_{i=1}^{L} Y_{ni}^2}{\sum_{i=1}^{L} Q_{ni}} \quad ,$$

The infit statistic for items (V_i) is calculated as:

$$V_i = \frac{\sum_{n=1}^{N} Y_{ni}^2}{\sum_{n=1}^{N} Q_{ni}} \quad ,$$

where Y^2_{ni} represents score residuals for items and Q_{ni} is an estimate of response variances (statistical information):

$$Q_{ni} = P_{ni}(1 - P_{ni}) \, ,$$

with P defined as the difficulty (p-value) for an item.

The expected value for these mean squares is 1.00, and various guidelines or "rules of thumb" have been proposed for identifying unacceptable departures from this expectation. Values of infit and outfit statistics that do not match the model-expected value of about 1.00 indicate that a facet may be influenced by construct-irrelevant factors. Essentially, these fit statistics provide an index of compatibility between the Rasch model and empirical data. Because Rasch models are probabilistic, some variation is expected. Overly determined or Guttman-like responses result in low values of fit statistics, and responses that are noisy and haphazard result in high values of fit statistics. Engelhard (2009) describes an acceptable range of infit and outfit statistics of about 0.80 to 1.20. Values that are lower than 0.80 suggest possible dependencies among responses, and values that are higher than about 1.20 suggest noisy responses; extreme values in both directions warrant further investigation.

Three-facet partial credit model. A three-facet version of the PC model that includes a parameter for subgroup membership (μ_m) can be written as:

$$ln\,[P_{nimk} / P_{nimk-1}] = \theta_n - \delta_i - \mu_m - \tau_{ik},$$

where

P_{nimk} = probability of person n on item i in subgroup m scoring k;
P_{nimk-1} = probability of person n on item i for subgroup m scoring $k - 1$;
θ_n = location of person n on latent variable;
δ_i = difficulty of item i;
μ_m = mean locations of subgroup m; and
τ_{ik} = difficulty of moving from category k–1 to k within item i.

This model includes three facets: person (θ_n), item (δ_i), and subgroup (μ_m). Three subgroup categories are included in the analyses: gender, race/ethnicity, and best language.

Analyses of DIF include an examination of item calibration differences on the logit scale between subgroups, as well as item-difficulty plots that provide graphical displays of the within-subgroup item calibrations. DIF maps are presented for aiding the substantive interpretation of item calibrations for selected subgroups. Within the context of Rasch measurement theory, differences in item calibrations on the logit scale between subgroups of students can be interpreted as effect sizes, as suggested by several researchers (Draba, 1977; Wright, Mead, & Draba, 1976). This approach for interpreting DIF has been used by Randall, Cheong, and Engelhard (2011) within the context of explanatory IRT modeling, and by Cheong (2006) for an analysis of school context effects on differential item functioning using hierarchical generalized linear models.

Residual Analyses, Person and Group Response Functions Based on Two-Facet Partial Credit Model. This section focuses on the analyses of residuals obtained after estimating the parameters from the PC model, as well as an examination of person fit statistics: outfit (unstandardized and standardized), slope parameters for persons and subgroups. Although slope parameters are not usually included as a parameter in Rasch models, the analyses described here view the slope (discrimination) parameter as a potentially useful tool for interpreting within-person and within-subgroup variability. Engelhard and Perkins (2011) have used this person slope parameter to aid in substantive interpretations of subgroup and person response functions. The slope parameter may represent differences in dispersion and units between persons and subgroups (Humphry, 2010).

Results

Two-facet partial credit model. The variable map shown in Figure 1 presents a graphical display of the spread of student measures (writing achievement), item locations (difficulty), and the location of the thresholds for the rating scale categories, all on the same logit scale. The Facets computer program (Linacre, 2007) was used to calibrate these facets. The first column shows the logit scale. The second column presents the student measures of writing achievement from the SAT-W. Higher-scoring students appear at the top of the column, and lower-scoring students appear at the bottom. Each asterisk represents 146 students, and a period represents one student. The student achievement measures range from −5.19 logits to 6.82 logits (M = 0.61, SD = 1.12, N = 19,341). The third column shows the item difficulty measures on the logit scale, with item difficulty ranging from −2.92 logits to 3.26 logits (M = 0.00, SD = 1.32, N = 51). Difficult items are located near the top of the column, and easier items are located closer to the bottom. As can be seen in Figure 1, the rating scale structure is comparable across both ratings of the essay.

Table 2 provides summary statistics from Facets analyses for the students, items, and student subgroups examined in this study. Items and student subgroups are centered at zero (mean set to zero), and only the average location of the student facet is allowed to vary. The overall differences between students (θ), items (δ), and each of the subgroups (gender, race/ethnicity, and best language) are significant ($p < 0.05$), with high reliabilities of separation ($Rel_{\theta} = 0.89$; $Rel_{\delta} > 0.99$; $Rel_{Gender} = 0.99$; $Rel_{Race/Ethnicity} > 0.99$; $Rel_{Best\ Language} > 0.99$). The reliability of separation statistic for persons from Facets is comparable to Cronbach's coefficient alpha. For other facets, the reliability of separation statistic describes the spread, or differences, between elements within a facet. The significant separation statistics for these students, items, and subgroups indicate a spread of the elements within each of the facets across the latent variable (writing achievement). Good fit to the model is evident for each of these facets, with mean infit and outfit statistics near their expected values of 1.00, with standard deviations around 0.20 (Engelhard, 2009). Acceptable model-data fit suggests that the many-facet Rasch (MFR) model is functioning as intended for these data.

Figure 2 shows the spread of item measures (difficulty) according to item subsets. The figure also demonstrates good targeting and alignment between the location and spread of item measures and the person measures (writing achievement). Each of the three item subsets has items located across a wide range of locations on the logit scale, with the Usage subset (U) demonstrating the largest spread. Similarly, Figure 3 shows the spread of student achievement by students in terms of their subgroup classifications. Small differences are evident for the gender subgroups, with females (M = 0.03) located only slightly higher on the logit scale than males (M = −0.03). For the racial/ethnic subgroups, a wide range of locations is evident with the Asian, Asian American, or Pacific Islander group located highest on the logit scale (M = 0.36), and the black or African American subgroup located lowest (M = −0.23). In terms of language groups, the English only group is located highest on the logit scale (M = 0.19), and the another language group is located lowest on the logit scale (M = −2.50).

A summary of item outfit statistics for each of the subgroups of interest is provided in Table 3. As described earlier, outfit statistics summarize residuals between observed responses and those that are expected based on the fitted model. These standard unweighted mean square statistics are sensitive to unexpected and unusual responses across the locations of items and persons. Outfit statistics summarize model-data fit, and they can be used to indicate individuals or groups of individual responses for whom items are functioning differently than expected by the model. Outfit statistics for the combined sample of students indicate good fit

to the model across all three item subsets, although Subset 1 (sentence correction items) and Subset 3 (revision-in context items) appear to have slightly less variation than expected: (outfit = 0.93, SD = 0.13) for Subset 1 and outfit = 0.98, SD = 0.15) for Subset 3. This may be related to the nature of the item subsets; the items in the revision-in-context subset are all related to a single passage of text. Overall, the analyses reported in Table 3 indicate very good fit for the SAT-W data to the model. There do not appear to be any item subsets that are functioning in an unexpected way across the subgroups of persons (gender, race/ethnicity, and best language).

There do not appear to be any item subsets that are functioning in an unexpected way across the subgroups of persons (gender, race/ethnicity, and best language).

In terms of gender, outfit statistics for the three item subsets appear comparable for males and females, although female students have slightly higher standard deviations for the usage items subset and the revision-in-context items subset than do male students. Outfit statistics are slightly more varied across race/ethnicity and best language groups. In terms of race/ethnicity, the highest values of outfit statistics, which indicate more variability in responses than expected by the model, were observed for black or African American and Hispanic students for the usage and revision-in-context item subsets. The lowest values were observed for the Asian subgroup of students on sentence correction items (M = 0.90, SD = 0.17). For the language groups, outfit statistics were highest overall for the another language group across all three item subsets.

Three-facet partial credit model. In order to examine DIF for the SAT-W items, item difficulties were calibrated using the Facets computer program (Linacre, 2007) separately for each subgroup of students based on gender, race/ethnicity, and best language, and the item calibrations were compared. As can be seen in Figure 4, the item difficulties are quite similar between the selected pairs of student subgroups. This suggests that the item difficulties are quite similar in value for all of the subgroups with the exception of the another best language (ABL) subgroup. The high correlations between item difficulties for the gender and race/ethnicity subgroups are likely a reflection of the fact that SAT-W items are screened for DIF prior to their operational use. The lower correlation found between the item difficulties for the EBL and ABL subgroups is likely a reflection of the fact that the SAT-W is not primarily designed for use by students whose best language is not English.

In order to further conceptualize DIF as a continuous variable, bar charts that are similar in appearance to variable maps were created to visually examine item calibration differences between selected subgroups of students. Two bar charts of logit differences are shown in Figure 5 and Figure 6 in order to illustrate the smallest and largest differences in item difficulty calibrations between student subgroups. The horizontal bars in Figures 5 and 6 reflect the magnitude and direction of the differences between item calibrations for the comparison groups. The difference values were calculated as the logit scale calibration of item difficulties for the focal groups (males and ABL) minus the calibration for the reference groups (females and EBL), such that positive values indicate that the reference group tends to score higher than the focal group. In Figure 5, the reference group for gender is male students, and

… item difficulties are quite similar between the selected pairs of student subgroups, with R^2 correlations above 0.92 for all comparisons except between the English best language (EBL) and another best language (ABL) subgroups ($R^2 = 0.82$).

the reference group in Figure 6 is the ABL students. The subset classification and item ID number for each SAT-W item is indicated on the DIF maps, with SC used to represent the sentence correction subset, U for the usage subset, RIC for the revision-in-context subset, and Ratings used to represent the two separate ratings for the essay item. The vertical alignment of the items shows the sequence in which students responded to each item in the test booklet. These figures provide a useful display for visualizing DIF between subgroups of students in terms of item characteristics. One rule of thumb for interpreting the substantive significance of the differences is to explore item differences that exceed an absolute value of .50 logits (Draba, 1977; Wright, Mead, & Draba, 1976). DIF maps for the racial/ethnic subgroups are not presented here.

Figure 5 illustrates DIF in terms of gender subgroups. As can be seen in this figure, DIF appears to vary across item subsets, although the magnitude of the gender differences is generally small. The directionality of both the sentence correction (SC) and usage subsets (U) are fairly evenly split between gender groups. Of the 25 SC items, females appear to have higher scores on 13 items; logit differences between these items range from –0.30 to +0.33 logits for the two gender groups. The 18 U items, whose logit differences between gender groups range from –0.54 to 0.33 logits are evenly split between males and females. In contrast, male students score higher on all but one revision in context (RIC) item (range of differences: –0.26 to 0.09 logits), and female students have higher scores on both essay ratings.

The magnitude and directionality of patterns for DIF shown in Figure 6 are somewhat different from Figure 5. Of the 25 SC items, the ABL group tends to score higher on 14 items, while the EBL group scores higher on 11 items. Logit differences for this item subset range from –0.32 to 0.29 logits. The U item subset, whose differences range from –0.34 to 0.20 logits, shows a similar pattern, with the ABL group scoring higher on 10 items, and the EBL group scoring higher on eight items. In terms of the RIC items, the EBL group scores higher on all but one item (range of differences: –0.59 to 0.32 logits). As expected based on previous research, the EBL group has higher scores on both essay ratings.

Residual Analyses, Person, and Group Response Functions Based on Two-Facet Partial Credit Model

Are students responding to SAT-W items as expected by the model? In this section, findings related to person fit statistics (standardized and unstandardized outfit, and slope parameter) and group response functions are examined for person subgroups across the total set of SAT-W items and as they relate to item subsets.

Person Fit. Table 4 shows the results from the Facets analyses for person fit. An examination of Table 4 reveals interesting findings related to the Rasch-based fit statistics for this model. The expected value of outfit statistics is 1.00, with a standard deviation of around 0.20, and the expected value of the standardized versions of these statistics (infit z and outfit z) is 0.00, with a standard deviation of 1.00 when good model-data fit is observed (Engelhard, 2009). The infit statistics as well as their standardized values indicate appropriate fit to the model for these data (M = 1.00, SD = 0.24; M = 0.03, SD = 1.07, respectively). In contrast, although the outfit statistics approximate their expected value of 1.00, their standard deviations are more than twice as large as expected by the model (M = 0.99, SD = 0.45). The largest value of the outfit standard deviation is observed for the ABL subgroup (M = 1.35, SD = 0.70). These high standard deviations suggest that the person subgroups are not responding as expected to the overall set of SAT-W items. The mean slope parameters across subgroups tend to match the expected value of 1.00, with the exception of the students for whom a language other than English is their best language (M = 0.76, SD = 0.35).

A similar story emerges when item subsets are examined separately. Tables 5 to 7 show person fit statistics for the sentence correction (SC), usage (U), and revision-in-context (RIC) items. As was the case for the total set of SAT-W items, an examination of person fit within each of these item subsets reveals high standard deviations for the outfit statistic, with the largest value found for the ABL subgroup across all three item subsets.

Person Response Functions. In order to illustrate person response functions associated with different values of fit statistics and person slopes, three individual students were selected who had theta (θ) estimates of 0.50 logits, but who varied in terms of their outfit statistics across the total set of SAT-W items. These three students are shown in Figure 7. Person 14025 in Panel A has a person response function with the steepest slope (slope = 1.43, outfit = .65), Person 18200 in Panel B has a slope close to 1.00 (slope = .97, outfit = 1.03), and Person 12313 in Panel C has the smallest slope with the worst fit (slope = 43, outfit = 1.77).

Figure 8 shows a way to examine the residual differences between expected responses and observed responses using residual plots for three other students. These plots show the residuals, or difference, between observed and expected responses based on the model, for each person's responses across all 51 SAT-W items. As illustrated in this figure, persons with high mean outfit statistics ("noisy" response patterns — Panel A), tend to have responses that vary from model expectations. Likewise, limited variation in responses is observed for persons with low outfit statistics ("muted" response patterns — Panel C). The residual plot shown in Panel B for a person who had an expected response pattern demonstrates that the Rasch model expects some variation in responses.

Group Response Functions. Figure 9 shows group response functions for selected student subgroups across the total set of SAT-W items. Group response functions are plots of the mean theta (θ) and slope estimates for specified subgroups of students. Group response functions can be used to examine differential group functioning (DGF). If the requirements for invariant measurement are met by a set of data, these functions will have comparable slopes

… the comparisons of group response functions for the race/ethnicity and best-language subgroups show relatively more variation than those for the two gender subgroups across all three item subsets.

and locations, and they will overlap. As seen in Figure 9, differences are observed across gender, race/ethnicity, and best-language subgroups on these items. As may be expected based on person fit indexes, larger variation is observed across the race/ethnicity and best-language subgroups than between the gender groups. Variation in response functions across student subgroups suggests that these groups are not responding as expected to the SAT-W items based on the model.

In order to examine item characteristics as a possible explanatory variable for differences in subgroup response patterns, group response patterns were compared across the three multiple-choice item subsets. Figures 10 to 12 illustrate group response functions for selected student subgroups within the sentence correction (SC), usage (U), and revision-in-context (RIC) item subsets; these figures correspond with the group measures and slopes given in Tables 5 to 7. As was observed for the total set of items, the comparisons of group response functions for the race/ethnicity and best-language subgroups show relatively more variation than those for the two gender subgroups across all three item subsets. An examination of these three figures indicates that the shape and location of each group vary across item subsets. It is interesting to note that group response functions related to items in the RIC subset show very little variation across student subgroups (because of nonindependence of the RIC items based on passage-based corrections of a paragraph), while more variation is observed in the SC and U subsets.

Summary and Discussion

The purpose of this study was to explore differential item functioning (DIF), differential person functioning (DPF), and differential group functioning (DGF) within the context of a large-scale writing assessment. The view of validity as a "function not only of the items or stimulus conditions, but also of the persons responding as well as the context of the assessment" (Messick, 1995, p. 741) lays the foundation for an examination of student writing achievement in terms of person and item characteristics. Furthermore, the role of communicative competence in career and college success establishes the need for a complete, contextualized view of score meaning on high-stakes assessments.

The SAT-W is designed to measure a variety of writing process skills related to both meaning (e.g., content and style) and mechanics (e.g., conventions and sentence formation) related to effective writing. As stated by Kobrin and Kimmel (2006), the guidelines for the development of the SAT-W are as follows:

- It should be accessible to the general test-taking population, including students for whom English is not a first or best language.

- It should be relevant to a wide range of fields and interests, and neither require specialized knowledge nor give an advantage to students who have completed a specific course of study.
- It should engage high-school-age students while stimulating critical reflection about important topics.
- It should be free of figurative or technical language or specific literary references.
- It should give students the opportunity to use a broad spectrum of experiences, learning, and ideas to support their point of view. (p. 2)

This study illustrates methodological tools for detecting DIF, DPF, and DGF based on an invariant measurement framework. The first research question is related to DIF for student subgroups based on gender, race/ethnicity, and best language, and the second question is related to DPF and DGF across item subsets with varying characteristics. Two item-response theory models based on Rasch measurement theory were used to examine the requirements of invariant measurement using data from an administration of the SAT-W, and various visual displays were used to examine and further illustrate findings from analyses. Good model-data fit for gender and race/ethnicity subgroups was found across item subsets. Variance in item and person functioning related to best-language subgroups suggested that some SAT-W items might be functioning differently for students whose best language is something other than English. This finding may be due to specific issues related to the students' particular language that are not reflected in the efforts to create an assessment for a general test-taking population. Students for whom English is not their best language may have also received differential opportunity to learn the English content included on the SAT-W. This study illustrates an additional set of analyses that can shed light on subgroup differences related to best language for the SAT-W.

A major strength of this study is that it uses data from a high-profile, large-scale assessment of writing that "defines" writing for college admission and for numerous students around the world. Because writing assessments such as the SAT-W combine high stakes with human judgment, their operational scores must be critically evaluated in order to ensure valid and fair opportunities for achievement across subgroups of students. Achievement differences set forth a challenge for researchers to develop assessments that are fair for all examinees. As stated in Standard 3.5 of the 1985 Standards for Educational and Psychological Testing:

> When selecting the type and content of items for tests and inventories, test developers should consider the content and type in relation to cultural backgrounds and prior experiences of the variety of ethnic, cultural, age, and gender groups represented in the intended population of test takers. (AERA/APA/NCME, 1985, p. 26)

Because writing assessments such as the SAT-W combine high stakes with human judgment, their operational scores must be critically evaluated in order to ensure valid and fair opportunities for achievement across subgroups of students.

The measurement community is challenged by this standard to ensure that the scores students receive on assessments are not influenced by characteristics unrelated to the construct of interest, such as gender, race/ethnicity, or best language. These principles are included in the current *Test Standards* (1999), and it is very likely that they will also appear in the next generation of revisions of this document.

This study employed methodological tools related to DIF and DPF that have been used to examine person responses in the context of large-scale assessments. However, the major focus of this study was on the use of explanatory models to detect systematic relationships between student and item characteristics and achievement differences (De Boeck & Wilson, 2004). A conceptual framework based on invariant measurement was presented, and models based on this framework were selected to illustrate the requirements of invariant measurement using data from the SAT-W. A key issue underlying this study was the interpretation of model-data misfit as a function of item characteristics, interpretation of items by persons, and interaction with contextual features that may influence responses. Model-data misfit can imply problems related to items themselves (DIF) or with student subgroup or individual interpretations of items (DGF and DPF), when groups or individuals do not respond to test items as expected and intended by the test developers. Methods for detecting DGF and DPF are promising for identifying individuals and groups for detailed qualitative interpretation, which may reveal differential opportunity to learn and other contextual factors related to unexpected responses.

One potential weakness of this study is related to the data used to illustrate these DIF and DPF methodologies. Because the data are from an operational administration of the SAT, the items used in this analysis have been prescreened for DIF by the College Board. However, because the purpose of this study is to illustrate explanatory models rather than to conduct bias or sensitivity reviews, data from the SAT-W provide an authentic and useful context in which to examine these issues. Future research should probe differential group and person functioning with specifically designed scales that are instructionally sensitive in order to detect differences in opportunity to learn. Based on a contextualized view of validity (Messick, 1995), and in light of this study's findings, more work is needed to build up a body of descriptive research on DIF, DPF, and DGF in order to clarify the construct of writing as a function of items, persons, and context. Research that examines and addresses score meaning can be supported with large-scale secondary data analyses that provide opportunities for authentic investigation of issues related to differential person and group functioning.

Similar to previous research on subgroup differences in writing achievement, this study also found fairly consistent subgroup differences with higher writing achievement for whites, females, and individuals for whom English is their best language. The current study extends this research by drilling down more deeply into interactions among item subsets and subgroup membership with individual and subgroup-level analyses of person fit. The focus on differential item and person functioning in regard to item subsets and person subgroups holds promise to add to the literature in these areas. Future research should continue to explore item, person, and subgroup differences related to writing achievement as measured by the SAT-W.

References

AERA, APA, & NCME (1985). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D. A. (1982). An index of person separation in latent trait theory, the traditional KR.20 indices and the Guttman scale response pattern. *Education Research and Perspectives, 9,* 95–104.

Behizadeh, N., & Engelhard, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing*, *16*, 189–211.

Bermúdez, A. B., & Prater, D. L. (1994). Examining the effects of gender and second language proficiency on hispanic writers' persuasive discourse. *Bilingual Research Journal*, *18*(3), 47–62.

Breland, H. M., Bonner, M. W., & Kubota, M. Y. (1995). *Factors in performance on brief, impromptu essay examinations*. New York: The College Board.

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework*. New York: College Board.

Breland, H., Kubota, M., Nickerson, K., Trapani, C., & Walker, M. (2004). *New SAT writing prompt study: Analyses of group impact and reliability* (College Board Research Report No. 2004-1). New York: The College Board.

Carlisle, R., & McKenna, E. (1990). Placement of ESL/EFL undergraduate writers in college-level writing programs. In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing* (pp. 197–209). Norwood, NJ: Ablex Publishing Corporation.

Cheong, Y. F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing*, *6*, 57–79.

Cole, N. S. (1997). Understanding gender differences and fair assessment in context. In W. W. Willingham (Ed.), *Gender and fair assessment* (pp. 157–184). Mahwah, NJ: Lawrence Erlbaum Associates.

Cole, N. S. (2000). *The ETS gender study: How females and males perform in educational settings*. Princeton, NJ: Educational Testing Service.

College Board. (2011). *Getting ready for the SAT*. New York: The College Board.

Common Core Standards Initiative. (2010). *Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects*. Prepared by the Council of Chief State School Officers and the National Governors Association.

Crane, E. W., Barrat, V. X., & Huang, M. (2011). *The relationship between English proficiency and content knowledge for English language learner students in grades 10 and 11 in Utah* (Issues & Answers Report, REL 2011–No. 110). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. Retrieved from http://ies.ed.gov/ncee/edlabs

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement,* 2nd edition (pp. 443–507). Washington, DC: American Council on Education.

Cumming, A. (2001). Learning to write in a second language: Two decades of research. *International Journal of English Studies*, *1*(2), 1–23.

Curley, W. E., & Schmitt, A. P. (1993). *Revising SAT-Verbal items to eliminate differential item functioning* (College Board Report No. 93-2). New York: College Entrance Examination Board.

De Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory item response models*. New York: Springer.

Draba, R. E. (1977). *The identification and interpretation of item bias* (Research Memorandum No. 25). Chicago: University of Chicago, MESA Psychometric Laboratory.

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. New York: Peter Lang.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341–349.

Engelhard, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken [Focus article]. *Measurement: Interdisciplinary Research and Perspectives 6*(3), 1–35.

Engelhard, G. (2009). Using item response theory and model data fit to conceptualize differential item functioning for students with disabilities. *Educational and Psychological Measurement*, *69*(4), 585–602.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.

Engelhard, G., Gordon, B., & Gabrielson, S. (1992). The influences of mode of discourse, experiential demand, and gender on the quality of student writing. *Research in the Teaching of English, 26*(3), 315–336.

Engelhard, G., Gordon, B., Walker, E. V., & Gabrielson, S. (1994). Writing tasks and gender: influences on writing quality of black and white students. *Journal of Educational Research, 87*, 197–209.

Engelhard, G., & Perkins, A. F. (2011). Person response functions and the definition of units in the social sciences. *Measurement: Interdisciplinary Research and Perspectives*, *9*, 40–45.

Fitzgerald, J. (2006). Multilingual writing in preschool through 12th grade: The last 15 years. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 337–354). New York: Guilford Press.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K. (1989). Principles and applications of item response theory. In R. L. Linn (Ed.), *Educational Measurement*, 3rd edition (pp. 147–200). New York: Macmillan.

Hamp-Lyons, L. (1991). Reconstructing "academic writing proficiency." In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts* (pp. 127–153). Norwood, NJ: Ablex Publishing.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly*, *37*(2), 275–301.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education*, *23*(4), 310–324.

Humphry, S. (2010). Modeling the effects of person group factors on discrimination. *Educational and Psychological Measurement*, *70*(2), 215–231.

Kane, M. T. (1992). An argument-based approach to validity. *Quantitative Methods in Psychology*, *112*(3), 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kobrin, J. L., & Kimmel, E. W. (2006). *Test development and technical information on the writing section of the SAT Reasoning Test* (Research Notes, RN-25). New York: The College Board.

Linacre, J. M. (2007). *A user's guide to FACETS: Rasch-model computer program*. Available online at www.winsteps.com

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linder & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). New York: Springer.

Mattern, K., Camara, W., & Kobrin, J. L. (2007). *SAT writing: An overview of research and psychometrics to date* (RN-32). New York: The College Board.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education*, *15*(4), 363–389.

Noeth, R. J., & Kobrin, J. L. 2007. *Writing changes in the nation's K–12 education system* (RN-37). New York: The College Board.

Oh, H., & Walker, M. E. (2006). *The effects of essay placement and prompt type on performance on the new SAT*. (College Board Research Report No. 2006-7). New York: The College Board.

Pomplun, M., Wright, D., Oleka, N., & Sudlow, M. (1992). *An analysis of English composition test essay prompts for differential difficulty*. New York: The College Board.

Randall, J., Cheong, Y. F., & Engelhard, G. (2011). Using explanatory IRT modeling to investigate context effects of differential item functioning for students with disabilities. *Educational and Psychological Measurement, 71*(1), 129–147.

Sathy, V., Barbuti, S., & Mattern, K. (2006). *The new SAT and trends in test performance* (Statistical Report No. 2006-1). New York: The College Board.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly, 27*(4), 657–677.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). Assessing the fit of item response theory models (pp. 683–718). In C. R. Rao and S. Sinharay (Eds.), *Psychometrics: Handbook of Statistics,* Volume 26. Amsterdam: Elsevier.

Tarone, E., Downing, B., Cohen, A., Gillette, S., Murie, R., & Dailey, B. (1993). The writing of southeast Asian-American students in secondary school and university. *Journal of Second Language Writing, 2*(2), 149–172.

Vann, R. J., Lorenz, F. O., & Meyer, D. M. (1991). Error gravity: Faculty response to errors in the written discourse of nonnative speakers of English. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 181–195). Norwood, NJ: Ablex Publishing Company.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 invitational conference on testing problems*. Princeton, NJ: Educational Testing Service.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Wright, B. D., Mead, R., & Draba, R. (1976). *Detecting and correcting test item bias with a logistic response model* (Research Memorandum No. 22). Chicago: University of Chicago, MESA Psychometric Laboratory.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233.

Table 1.

Demographic Information

	Subgroups	Total (*N* = 388,889)		5% Random Sample (*n* = 19,341)	
		N	%	*n*	%
Gender	Female	219,035	56.30%	10,978	56.80%
	Male	169,864	43.70%	8,363	43.20%
Race/Ethnicity	American Indian or Alaska Native	1,803	0.50%	94	0.50%
	Asian, Asian American, or Pacific Islander	39,874	10.30%	1,920	9.90%
	Black or African American	36,605	9.40%	1,857	9.60%
	Hispanic*	53,468	13.70%	2,632	13.60%
	White	235,736	60.60%	11,801	61.00%
	Other	9,810	2.50%	488	2.50%
	No Response	11,603	3.00%	549	2.80%
Best Language	English Only	322,011	82.80%	16,066	83.10%
	English and Another Language	55,968	14.40%	2,752	14.20%
	Another Language	5,753	1.50%	273	1.40%
	No Response	4,701	1.20%	228	1.20%
Mean SAT Writing Score (*SD*)		511.6 (105.1)		511.3 (104.7)	

*Note: Hispanic group includes Mexican, Mexican American, Puerto Rican, and Other Hispanic, Latino, or Latin American students.

Table 2.

Summary Statistics from Facets Analysis

			Subgroups		
	Students (n = 19,341)	**Items**	**Gender**	**Race/Ethnicity**	**Best Language**
Measures					
M	.61	.00	.00	.00	.00
SD	1.12	1.32	.04	.23	.23
Count	19,341	51	2	6	3
Infit					
M	1.00	.99	1.01	1.02	1.05
SD	.24	.09	.02	.01	.06
Outfit					
M	.99	.99	.99	1.01	1.12
SD	.45	.21	.02	.04	.21
Reliability of Separation	.89	> .99	.99	> .99	> .99
X^2 Statistic	149,327.1*	186,530.8*	133.5*	4,797.1*	787.5*
Degrees of Freedom	19,340	50	1	5	2

* $p < .05$

Note: Item statistics include both multiple-choice items and two ratings of the essays.

Table 3.

Summary of Outfit Statistics by Item Subsets for Selected Subgroups

	Total	Gender		Race/Ethnicity				Best Language		
Item Subsets	**(*n* = 19,341)**	**Male (*n* = 8,363)**	**Female (*n* = 10,978)**	**Asian, Asian American, or Pacific Islander (*n* = 1,920)**	**Black or African American (*n* = 1,857)**	**Hispanic* (*n* = 2,632)**	**White (*n* = 11,801)**	**English Only (*n* = 16,066)**	**English and Another Language (*n* = 2,752)**	**Another Language (*n* = 273)**
Subset 1: Sentence Correction (25 Items)										
M	.93	.93	.93	.90	.96	.96	.94	.93	.94	.98
SD	.13	.13	.13	.17	.12	.12	.14	.13	.13	.21
Subset 2: Usage (18 Items)										
M	1.05	1.05	1.05	.99	1.06	1.06	1.04	1.05	1.04	1.07
SD	.29	.26	.32	.20	.28	.27	.28	.29	.22	.28
Subset 3: Revision in Context (6 Items)										
M	.98	.98	.98	.99	1.06	1.02	.97	.98	1.00	1.12
SD	.15	.12	.17	.16	.21	.16	.12	.15	.16	.20

* Note: Hispanic group includes Mexican, Mexican American, Puerto Rican, and Other Hispanic, Latino, or Latin American students.

Table 4.

Summary of Person Measures, Outfit Statistics, Standardized Outfit Statistics, and Slopes by Subgroups

	Measure	Outfit	Standardized Outfit	Slope	*n*
	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	
Gender					
Female	0.92 (1.16)	0.97 (0.44)	0.00 (0.99)	1.00 (0.30)	10,978
Male	0.83 (1.16)	1.00 (0.46)	0.07 (1.02)	0.98 (0.32)	8,363
Race/Ethnicity					
Asian, Asian American, or Pacific Islander	1.25 (1.32)	1.00 (0.52)	0.13 (0.96)	0.96 (0.29)	1,920
Black or African American	0.23 (1.05)	1.05 (0.40)	0.15 (1.10)	0.96 (0.34)	1,857
*Hispanic**	0.36 (1.04)	1.06 (0.45)	0.20 (1.13)	0.95 (0.34)	2,632
White	1.04 (1.09)	0.96 (0.43)	-0.04 (0.96)	1.02 (0.29)	11,801
Best Language					
English Only	0.94 (1.14)	0.97 (0.43)	-0.01 (0.98)	1.01 (0.29)	16,066
English and Another Language	0.64 (1.17)	1.04 (0.46)	0.18 (1.05)	0.95 (0.32)	2,752
Another Language	0.17 (1.25)	1.35 (0.70)	0.85 (1.22)	0.76 (0.35)	273
TOTAL	0.88 (1.16)	0.99 (0.45)	0.03 (1.00)	0.99 (0.31)	19,341

*Note: Hispanic group includes Mexican, Mexican American, Puerto Rican, and Other Hispanic, Latino, or Latin American students.

Table 5.

Summary of Person Fit by Subgroups: Sentence Correction Items (25 Items)

	Measure	Outfit	Standardized Outfit	Slope	*n*
	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	
Gender					
Female	1.35 (1.42)	0.97 (0.57)	0.05 (0.85)	1.00 (0.36)	10,977
Male	1.25 (1.40)	0.98 (0.56)	0.07 (0.86)	0.99 (0.38)	8,361
Race/Ethnicity					
Asian, Asian American, or Pacific Islander	1.71 (1.51)	0.96 (0.58)	0.09 (0.80)	0.99 (0.34)	1,920
Black or African American	0.60 (1.30)	1.04 (0.50)	0.13 (0.96)	0.95 (0.43)	1,857
*Hispanic**	0.75 (1.26)	1.02 (0.48)	0.13 (0.91)	0.96 (0.40)	2,632
White	1.27 (1.47)	0.96 (0.55)	0.05 (0.80)	1.00 (0.34)	11,798
Best Language					
English Only	1.37 (1.64)	0.97 (0.57)	0.04 (0.86)	1.00 (0.362)	16,063
English and Another Language	1.04 (1.38)	1.00 (0.50)	0.12 (0.88)	0.97 (0.38)	2,752
Another Language	0.53 (1.42)	1.22 (0.66)	0.45 (0.98)	0.82 (0.44)	273
TOTAL	1.31 (1.41)	0.97 (0.56)	0.06 (0.85)	0.99 (0.37)	19,338

*Note: Hispanic group includes Mexican, Mexican American, Puerto Rican, and Other Hispanic, Latino, or Latin American students.

Table 6.

Summary of Person Fit by Subgroups: Usage Items (18 Items)

	Measure	Outfit	Standardized Outfit	Slope	*n*
	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	
Gender					
Female	0.57 (1.33)	1.01 (0.81)	0.14 (0.85)	1.01 (0.48)	8,346
Male	0.51 (1.33)	1.04 (0.84)	0.17 (0.86)	0.99 (0.48)	10,972
Race/Ethnicity					
Asian, Asian American, or Pacific Islander	0.89 (1.55)	1.12 (0.91)	0.29 (0.84)	0.92 (0.49)	1,920
Black or African American	-0.06 (1.25)	1.09 (0.92)	0.17 (0.94)	0.98 (0.52)	1,856
*Hispanic**	0.12 (1.25)	1.16 (0.98)	0.26 (0.97)	0.92 (0.54)	2,629
White	0.70 (1.26)	0.97 (0.72)	0.10 (0.80)	1.04 (0.45)	11,784
Best Language					
English Only	0.60 (1.31)	0.99 (0.76)	0.12 (0.83)	1.02 (0.46)	16,047
English and Another Language	0.32 (1.36)	1.14 (0.96)	0.26 (0.92)	0.92 (0.52)	2,749
Another Language	-0.13 (1.53)	1.66 (1.50)	0.77 (1.09)	0.61 (0.61)	272
TOTAL	0.55 (1.34)	1.02 (0.82)	0.15 (0.85)	1.00 (0.48)	19,296

*Note: Hispanic group includes Mexican, Mexican American, Puerto Rican, and Other Hispanic, Latino, or Latin American students.

Table 7.

Summary of Person Fit by Subgroups: Revision in Context Items (6 Items)

	Measure	Outfit	Standardized Outfit	Slope	*n*
	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	
Gender					
Female	0.44 (1.26)	0.95 (1.00)	0.04 (0.80)	1.09 (0.83)	7,765
Male	0.44 (1.26)	0.94 (0.99)	0.02 (0.79)	1.12 (0.83)	6,122
Race/Ethnicity					
Asian, Asian American, or Pacific Islander	0.46 (1.26)	0.94 (1.02)	0.02 (0.80)	1.11 (0.78)	1,466
Black or African American	0.41 (1.27)	0.91 (0.91)	0.01 (0.78)	1.12 (0.83)	1,342
*Hispanic**	0.41 (1.26)	0.93 (0.94)	0.02 (0.79)	1.12 (0.84)	2,093
White	0.45 (1.26)	0.96 (1.03)	0.04 (0.81)	1.10 (0.83)	8,124
Best Language					
English Only	0.43 (1.26)	0.95 (0.99)	0.03 (0.80)	1.10 (0.83)	11,312
English and Another Language	0.46 (1.26)	0.93 (0.99)	0.02 (0.80)	1.12 (0.82)	2,147
Another Language	0.45 (1.27)	0.94 (1.01)	0.03 (0.80)	1.13 (0.87)	234
TOTAL	0.44 (1.26)	0.95 (0.99)	0.03 (0.80)	1.10 (0.83)	13,887

*Note: Hispanic group includes Mexican, Mexican American, Puerto Rican, and Other Hispanic, Latino, or Latin American students.

Figure 1.

Variable map for items, persons, and essay ratings.

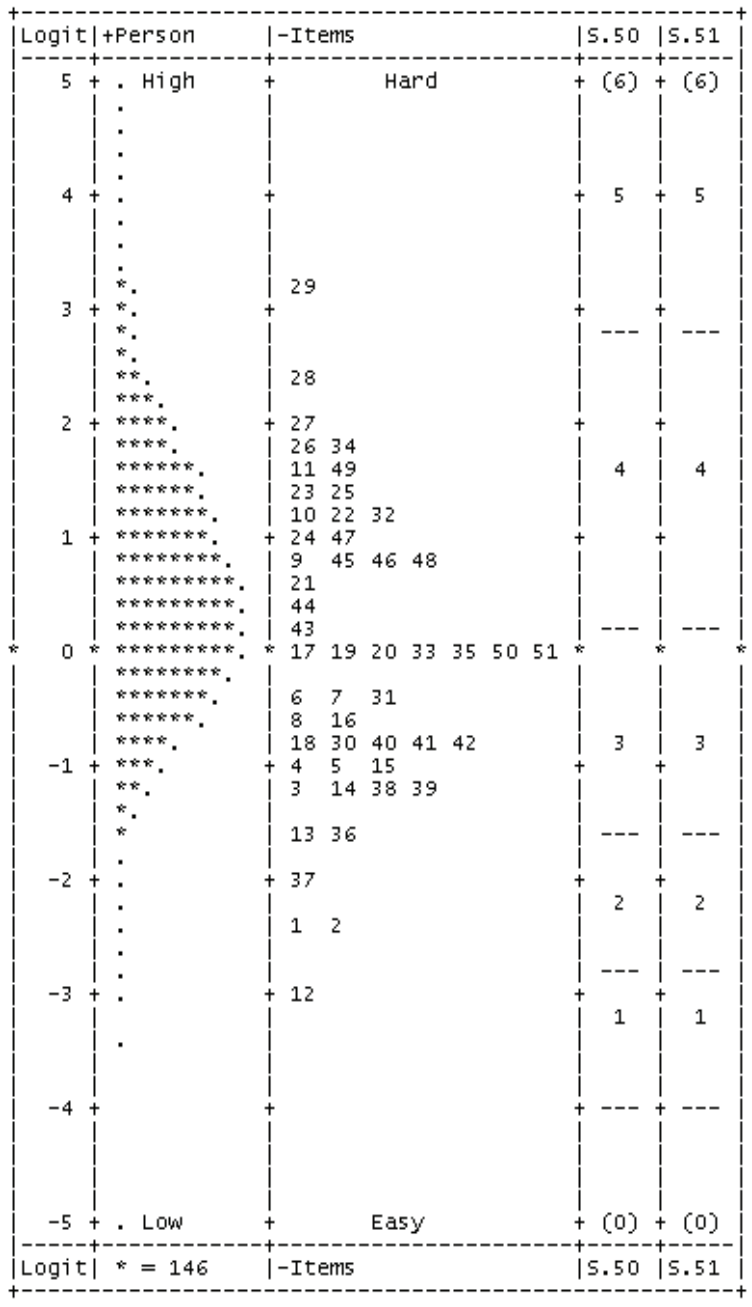
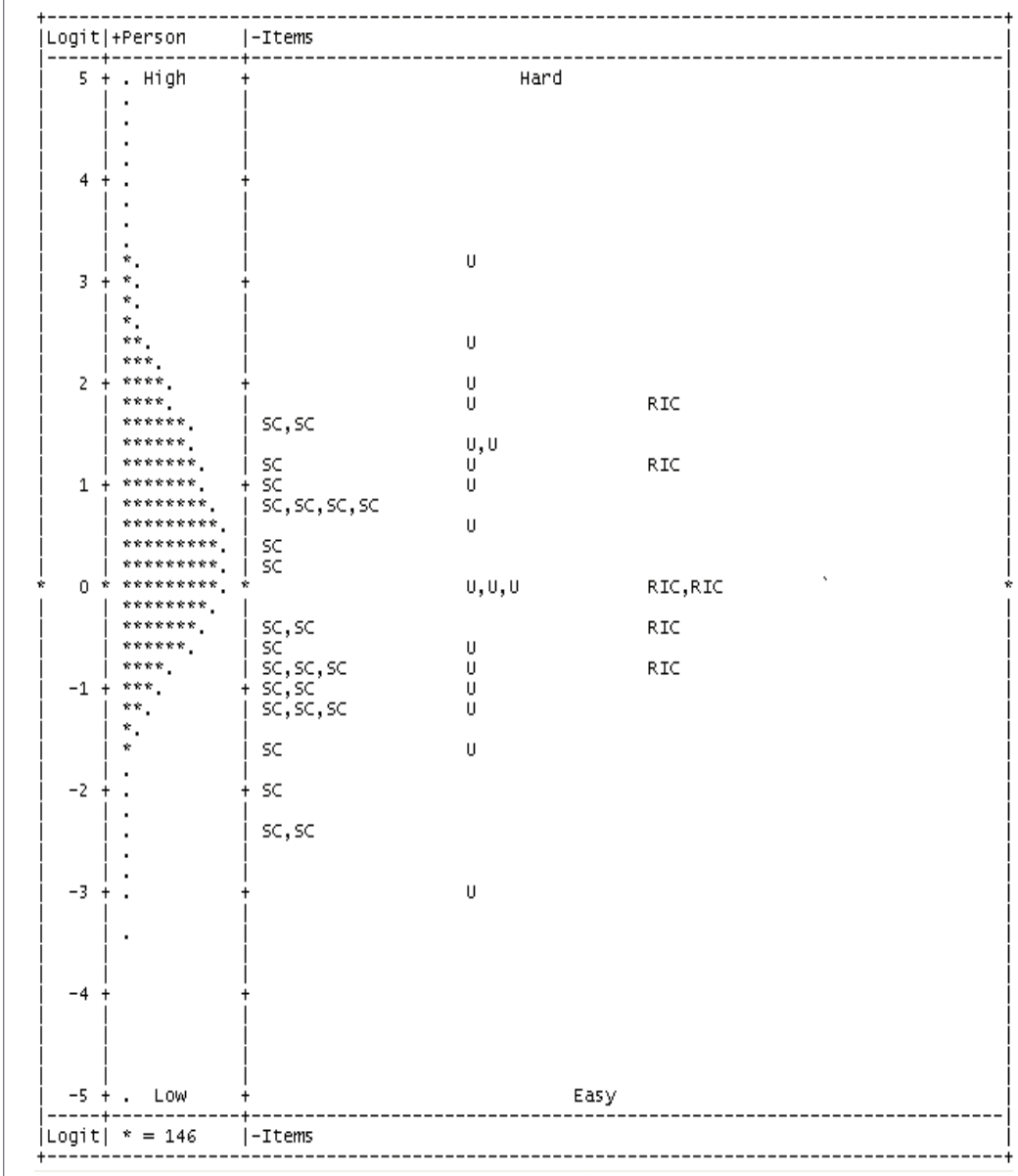


Figure 2.

Variable map for persons and items by category.



Item Subsets:
SC: Sentence Correction
U: Usage
RIC: Revision in Context

Figure 3.

Variable map for explanatory variables.

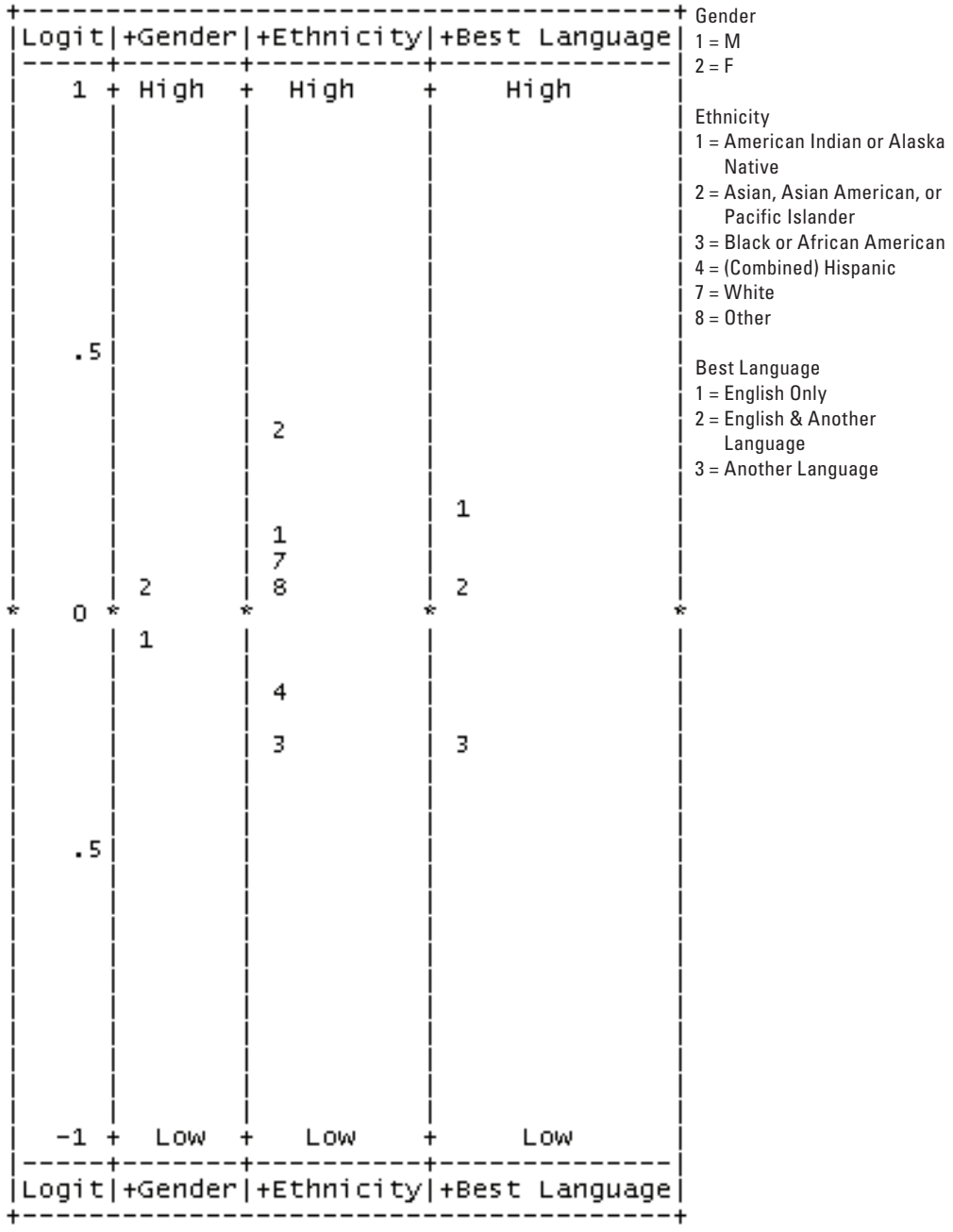


Figure 4.

Item calibrations for selected subgroup comparisons.

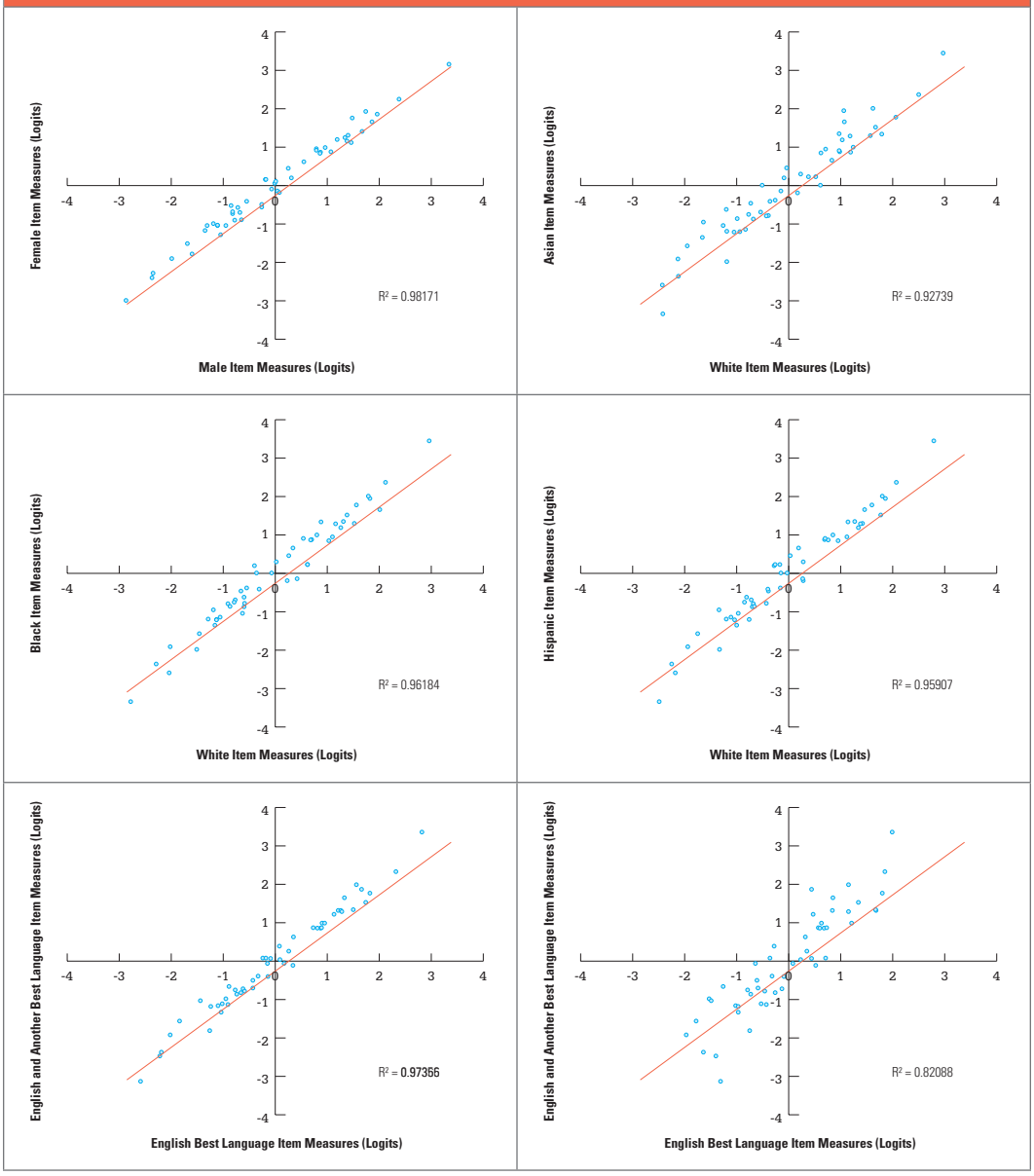
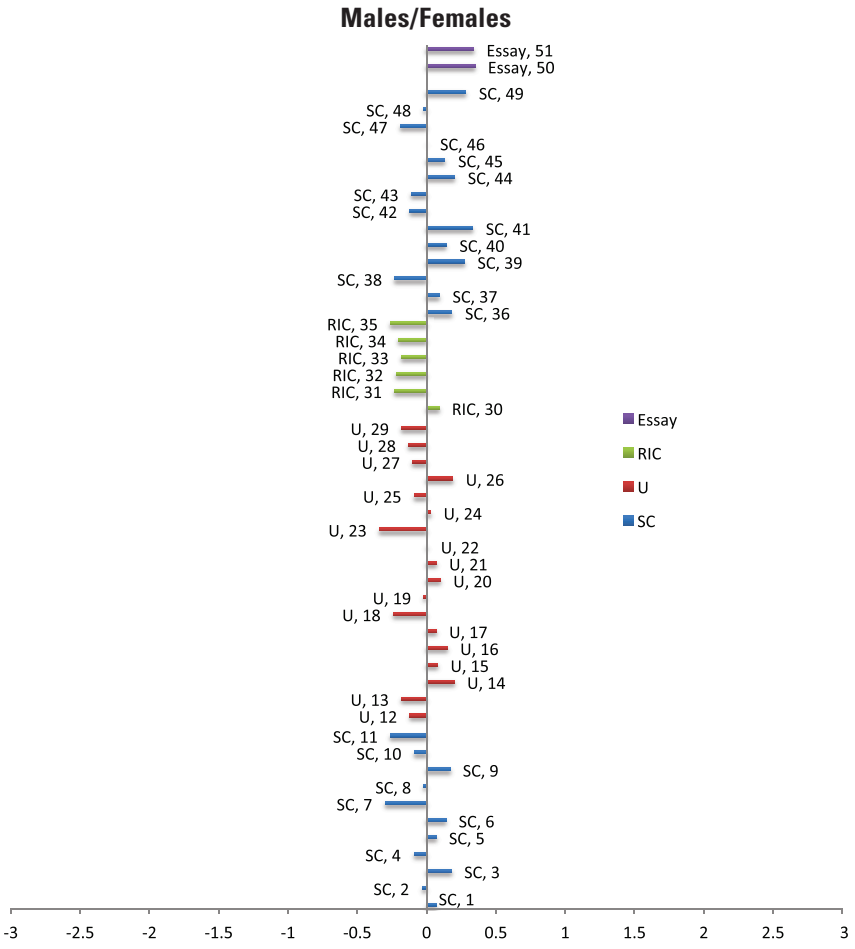


Figure 5.

DIF map for gender.

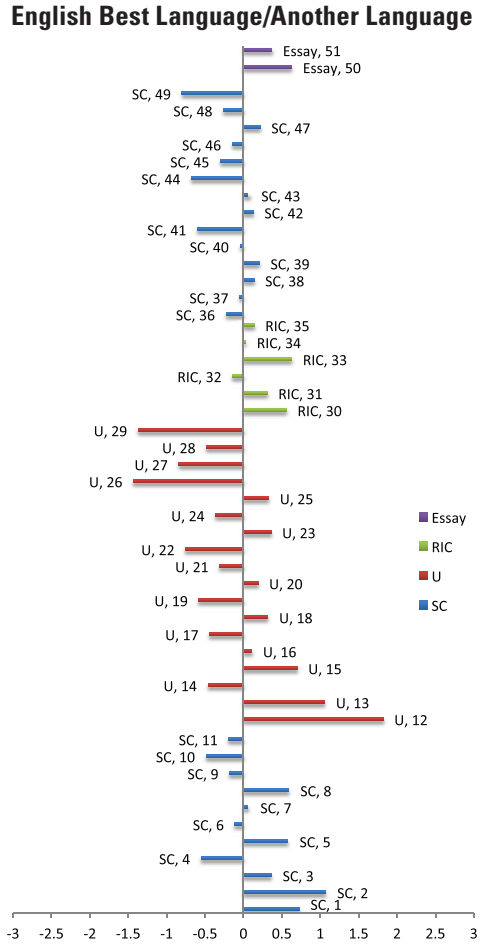


Item Subsets:

- SC: Sentence Correction
- U: Usage
- RIC: Revision in Context
- Ratings: Two Essay Ratings

Figure 6.

DIF map for best language.



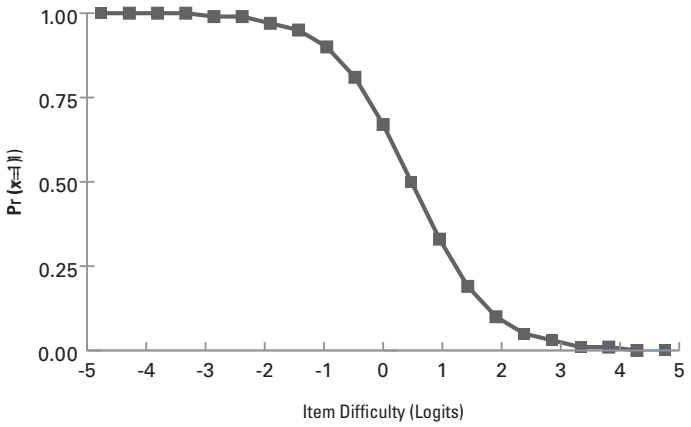
Item Subsets:

- SC: Sentence Correction
- U: Usage
- RIC: Revision in Context
- Ratings: Two Essay Ratings

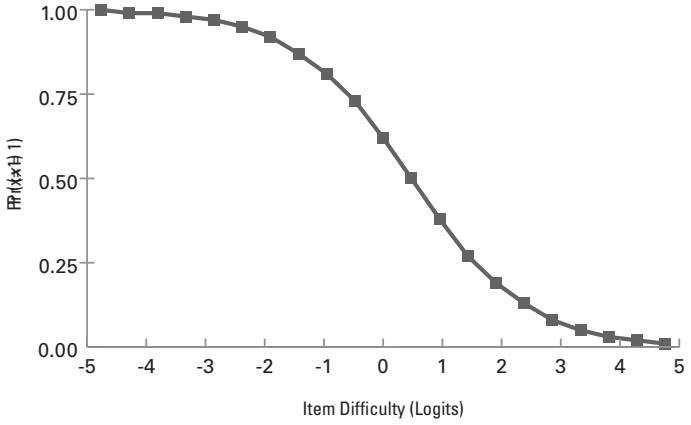
Figure 7.

Expected person response functions.

Panel A
Person 14025, Theta = 0.50 Logits, Outfit = .65, Slope = 1.43



Panel B
Person 18200, Theta = 0.50 Logits, Outfit = 1.03, Slope = .97



Panel C
Person 12313, Theta = 0.50 Logits, Outfit = 1.77, Slope = .43

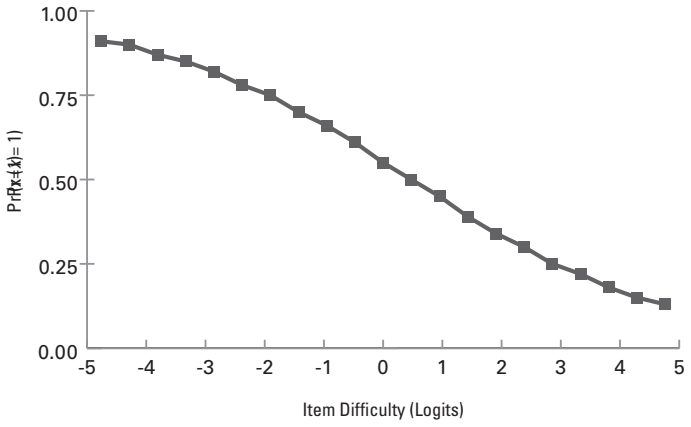
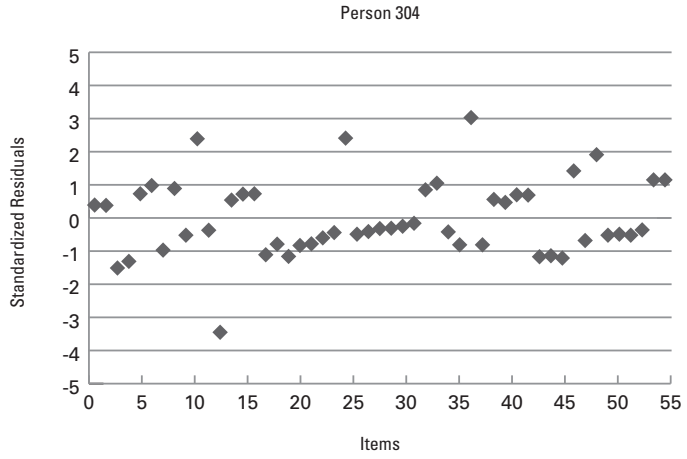


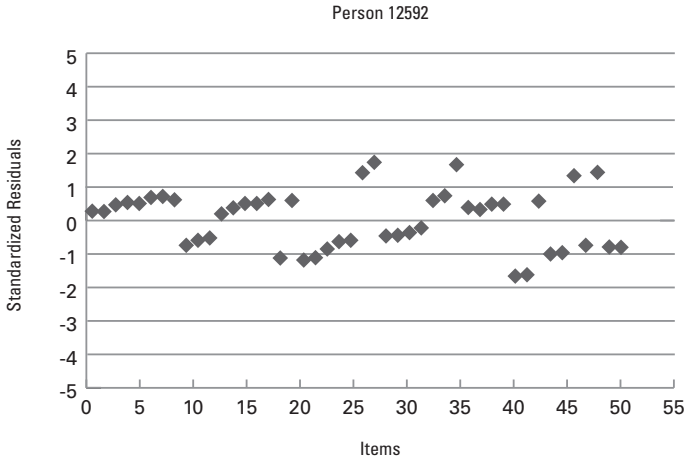
Figure 8.

Residual analyses for person response functions.

Panel A
Noisy Response Pattern [Outfit = 2.03]



Panel B
Expected Response Pattern [Outfit = 1.00]



Panel C
Muted Response Pattern [Outfit = 0.51]

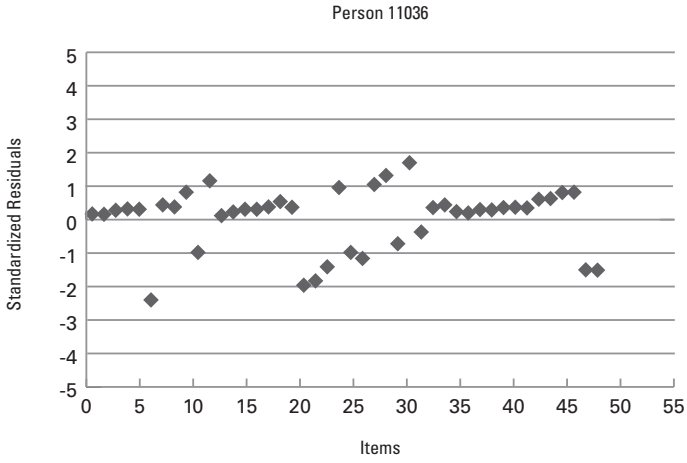
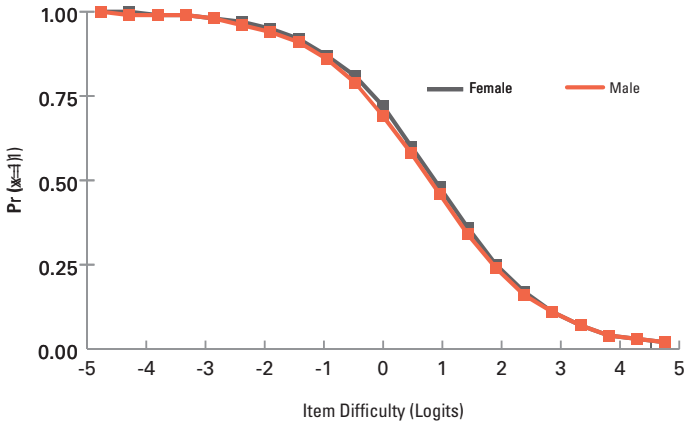


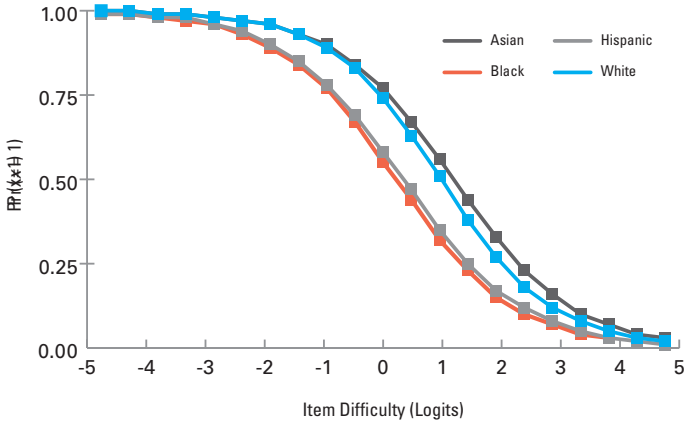
Figure 9.

Group response functions: Total set of items.

Gender



Race/Ethnicity



Best Language

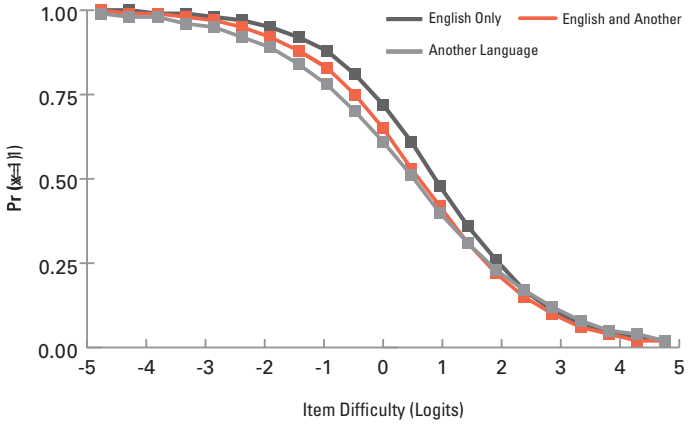
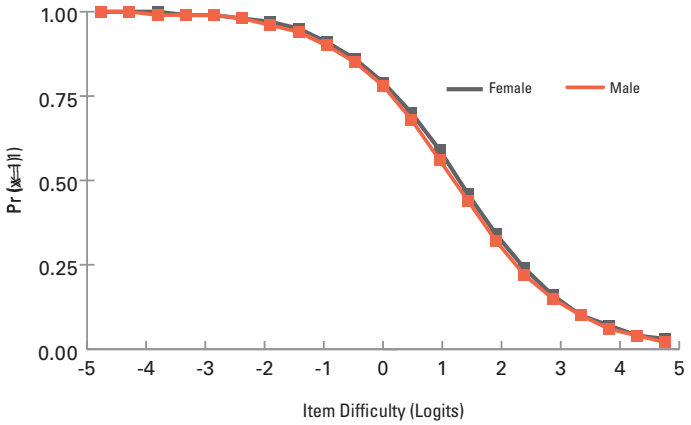


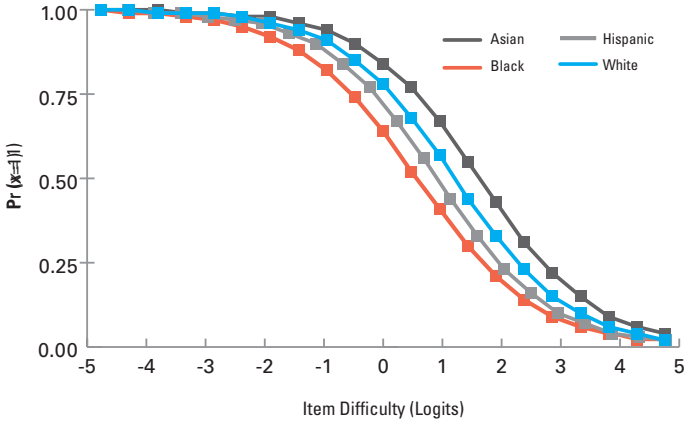
Figure 10.

Group response functions: Sentence correction items.

Gender



Race/Ethnicity



Best Language

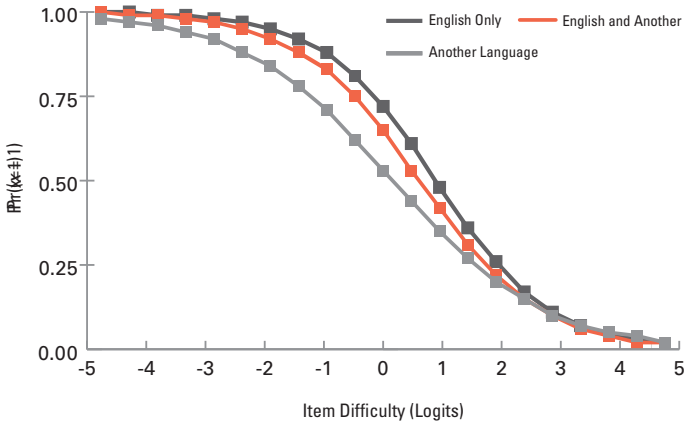
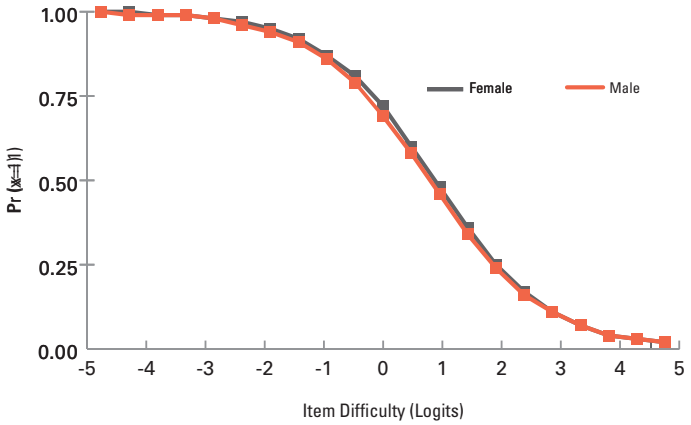


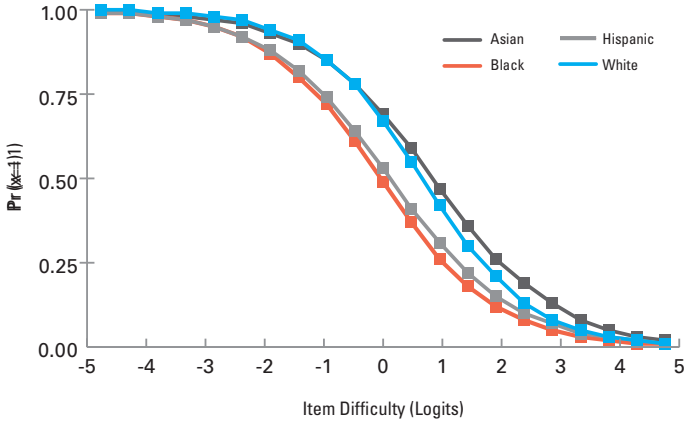
Figure 11.

Group response functions: Usage items.

Gender



Race/Ethnicity



Best Language

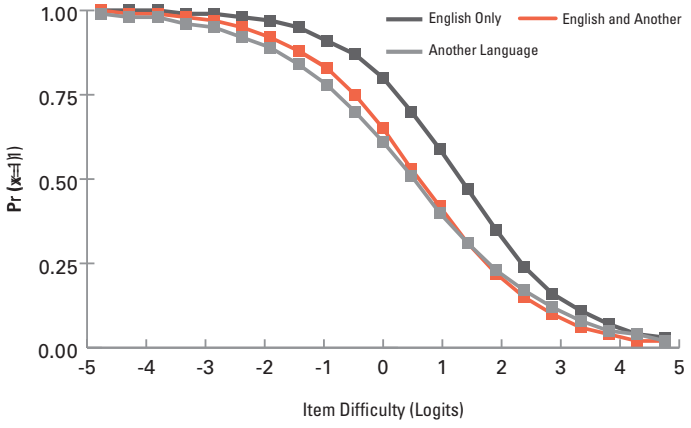
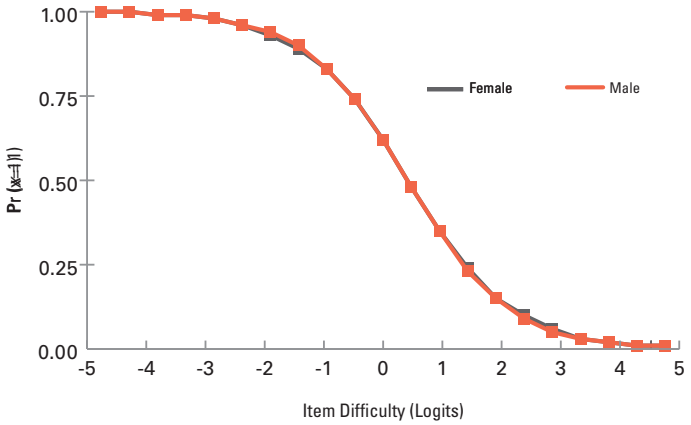


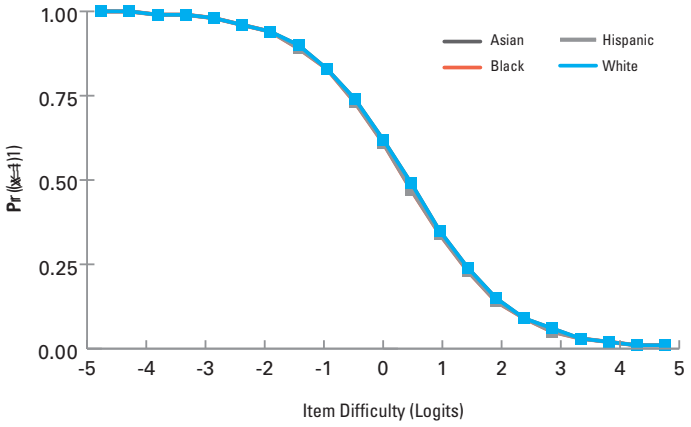
Figure 12.

Group response functions: Revision in context items.

Gender



Race/Ethnicity



Best Language

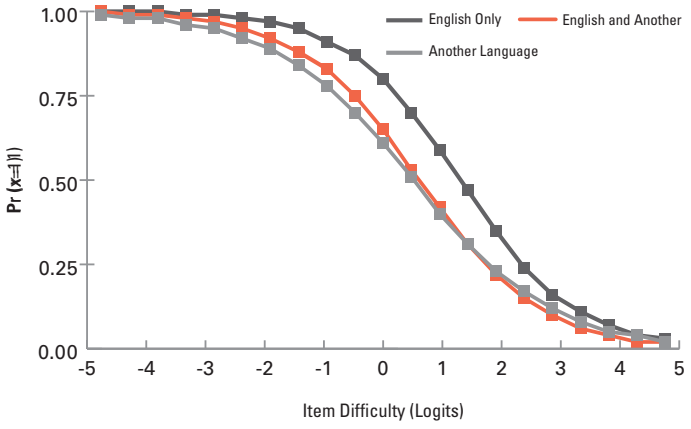


Table A1. Rubric for Essay

Group response functions: Revision in context items.

Score Point	1	2	3	4	5	6
Description	An essay in this category demonstrates *very little mastery*, and is severely flawed by ONE OR MORE of the weaknesses listed below.	An essay in this category demonstrates *little mastery*, and is flawed by ONE OR MORE of the weaknesses listed below.	An essay in this category demonstrates *developing mastery*, and is marked by ONE OR MORE of the weaknesses listed below.	An essay in this category demonstrates *adequate mastery*, although it will have lapses in quality. A typical essay:	An essay in this category demonstrates *reasonably consistent mastery*, although it will have occasional errors or lapses in quality. A typical essay:	An essay in this category demonstrates *clear and consistent mastery*, although it may have a few minor errors. A typical essay:
Features						
Development	Develops no viable point of view on the issue, or provides little or no evidence to support its position	Develops a point of view on the issue that is vague or seriously limited and demonstrates weak critical thinking, providing inappropriate or insufficient examples, reasons, or other evidence to support its position	Develops a point of view on the issue, demonstrating some critical thinking, but may do so inconsistently or use inadequate examples, reasons, or other evidence to support its position	Develops a point of view on the issue and demonstrates competent critical thinking, using adequate examples, reasons, and other evidence to support its position	Effectively develops a point of view on the issue and demonstrates strong critical thinking, generally using appropriate examples, reasons, and other evidence to support its position	Effectively and insightfully develops a point of view on the issue and demonstrates outstanding critical thinking, using clearly appropriate examples, reasons, and other evidence to support its position
Organization and Focus	Is disorganized or unfocused, resulting in a disjointed or incoherent essay	Is poorly organized and/or focused, or demonstrates serious problems with coherence or progression of ideas	Is limited in its organization or focus, or may demonstrate some lapses in coherence or progression of ideas	Is generally organized and focused, demonstrating some coherence and progression of ideas	Is well organized and focused, demonstrating coherence and progression of ideas	Is well organized and clearly focused, demonstrating clear coherence and a smooth progression of ideas
Use of Language	Displays fundamental errors in vocabulary	Displays very little facility in the use of language, using very limited vocabulary or incorrect word choice	Displays developing facility in the use of language, but sometimes uses weak vocabulary or inappropriate word choice	Exhibits adequate but inconsistent facility in the use of language, using generally appropriate vocabulary	Exhibits facility in the use of language, using appropriate vocabulary	Exhibits skillful use of language using a varied, accurate, and apt vocabulary
Sentence Variety	Demonstrates severe flaws in sentence structure	Demonstrates frequent problems in sentence structure	Lacks variety or demonstrates problems in sentence structure	Demonstrates some variety in sentence structure	Demonstrates variety in sentence structure	Demonstrates meaningful variety in sentence structure
Conventions	Contains pervasive errors in grammar, usage, or mechanics that persistently interfere with meaning	Contains errors in grammar, usage, and mechanics so serious that meaning is somewhat obscured	Contains an accumulation of errors in grammar, usage, and mechanics	Has some errors in grammar, usage, and mechanics	Is generally free of most errors in grammar, usage, and mechanics	Is free of most errors in grammar, usage, and mechanics

Note: Ratings of "0" are assigned to blank essays, off-topic essays, or illegible essays.

The Research department actively supports the College Board's mission by:

- Providing data-based solutions to important educational problems and questions
- Applying scientific procedures and research to inform our work
- Designing and evaluating improvements to current assessments and developing new assessments as well as educational tools to ensure the highest technical standards
- Analyzing and resolving critical issues for all programs, including AP®, SAT®, PSAT/NMSQT®
- Publishing findings and presenting our work at key scientific and education conferences
- Generating new knowledge and forward-thinking ideas with a highly trained and credentialed staff

Our work focuses on the following areas

Admission	Measurement
Alignment	Research
Evaluation	Trends
Fairness	Validity

Follow us online: research.collegeboard.org

CollegeBoard

CollegeBoard

12b 6991