College Board Research Report No. 2003-1

# Monitoring Faculty Consultant Performance in the Advanced Placement English Literature and Composition Program with a Many-Faceted Rasch Model

George Engelhard, Jr., and Carol M. Myford

College Board Research Report No. 2003-1
ETS RR-03-01

# Monitoring Faculty Consultant Performance in the Advanced Placement English Literature and Composition Program with a Many-Faceted Rasch Model

George Engelhard, Jr., and Carol M. Myford

College Entrance Examination Board, New York, 2003

George Engelhard, Jr., is a professor in the department of educational studies at Emory University.

Carol M. Myford is an associate professor of educational psychology in the College of Education at the University of Illinois at Chicago. She was formerly a senior research scientist in the Center for Measurement Models at Educational Testing Service.





Additional copies of this report (item #995947) may be obtained from College Board Publications, Box 886, New York, NY 10101-0886, 800 323-7155. The price is $15. Please include $4 for postage and handling.



Printed in the United States of America.

## *Acknowledgments*

We would like to acknowledge the helpful advice of Mike Linacre (University of Chicago) regarding the use of the FACETS computer program to analyze the data. Belita Gordon (University of Georgia) assisted us with the preparation of the literature review. The material contained herein is based on work supported by the Advanced Placement Research and Development Committee. Any opinions, findings, conclusions, and recommendations expressed herein are those of the authors and do not necessarily reflect the views of the College Board, Emory University, or the Educational Testing Service.

# Contents

*Tables*

*Figures*

# Abstract

The purpose of this study was to examine, describe, evaluate, and compare the rating behavior of faculty consultants who scored essays written for the Advanced Placement English Literature and Composition (AP® ELC) Exam. Data from the 1999 AP ELC Exam were analyzed using FACETS (Linacre, 1998) and SAS. The faculty consultants were not all interchangeable in terms of the level of severity they exercised. If students' ratings had been adjusted for severity differences, the AP grades of about 30 percent of the students would have been different from the one they received. Almost all the differences were one grade or less. Adjusting ratings for faculty consultant severity differences would not impact some student subgroups more than others.

**Keywords:** raters, rater effects, performance assessment, item response theory, Rasch measurement, FACETS, rater monitoring, quality control, Advanced Placement Program®

# Introduction

Researchers have detected variation in the level of severity that raters exercise when scoring essays in a variety of operational assessment systems that depend on ratings, including the Advanced Placement Program® (AP®) Examinations. Results from their studies indicate that rater severity differences are more pronounced in the scoring of some AP Exams than in the scoring of others (Braun, 1988; Braun and Wainer, 1989; Bridgeman, Morgan, and Wang, 1996; Longford, 1994a, 1994b; Morgan, 1998). The AP Program currently uses the Reader Management System to monitor rater behavior for some of its exams. However, there are several new and promising item response theory (IRT)–based approaches for evaluating the quality of ratings that have been developed over the past 10 years. The major purpose of this study was to determine whether one IRT-based approach, many-faceted Rasch measurement, could provide additional safeguards. The overall goal of implementing this approach is to help ensure that each student who takes an AP Exam receives an AP grade that is a fair and accurate measure of the student's achievement, regardless of the particular faculty consultants who happen to rate the student's essays.

In this study, we analyzed data from the 1999 AP English Literature and Composition (ELC) Examination. Some of the IRT-based approaches for analyzing rating data cannot be used with the AP ELC Exam because a single faculty consultant rates each student's essay. (The AP ELC Exam includes three free-response questions, and the student writes an essay for each question.) In terms of design constraints, faculty consultants are nested within the three free-response questions, and there is no overlap among faculty consultants. This incomplete block design (Ebel, 1951; Fleiss, 1981) characterizes a number of AP Examinations, posing a challenge for all IRT-based approaches to analyzing the data, since the data matrices resulting from this type of design are typically sparse, containing much missing data. In order to address this lack of connectivity among faculty consultants, the study explored a promising approach for calibrating faculty consultants in order to examine severity. Once the faculty consultants are connected, then standard many-faceted Rasch measurement computer programs like FACETS (Linacre, 1998) can be used to analyze the AP data. In addition to exploring faculty consultant severity, the study examined interactions between rater severity and selected student characteristics (i.e., gender, race/ethnicity, and best language) that may impact essay ratings and AP grades. These differential facet functioning analyses provided information regarding potential sources of bias in the ratings.

## *Purpose of the Study*

This study is an investigation of faculty consultant performance in the scoring of the free-response questions included on the 1999 AP English Literature and Composition Exam. The purpose of this study is to examine, describe, evaluate, and compare the rating behavior of individual faculty consultants. The intent of the study is to determine to what extent faculty consultants may be introducing construct-irrelevant variance into the assessment process. Specifically, this study addresses the following questions:

1. Do faculty consultants differ in the levels of severity they exercise when scoring students' essays written for Section II of the 1999 AP English Literature and Composition Exam? What is the best approach for calibrating faculty consultants?
2. Are there interactions between faculty consultant severity and extraneous student background characteristics (e.g., gender, race/ethnicity, and best language) that may impact essay ratings and grades on the 1999 AP English Literature and Composition Examination?
3. Do adjustments for faculty consultant severity have an impact on essay ratings and/or on AP grades?

4. Does faculty consultant severity differentially impact essay ratings and/or the AP grades for student subgroups based on student gender, race/ethnicity, or best language?

# Review of the Literature

## *Variation in Rater Severity*

Rater severity/leniency is the systematic assignment of lower or higher ratings than the average of ratings assigned by other raters. It has been identified as a rater effect and/or rater error in AP Examinations, including the AP English Literature and Composition (AP ELC) Exam, in studies of large-scale writing assessments, and in other performance assessment arenas.

Rater severity differences have been reported in four large-scale writing assessments. In both eighth grade writing (Engelhard, 1994) and high school writing (Gyagenda and Engelhard, 1998) significant differences in levels of severity exercised were found in spite of extensive rater training. Similar findings were reported by Du and Wright (1997) for students who produced two essays in grades six, eight, and ten. While these findings were based on a single scoring session, Fitzpatrick, Ercikan, Yen, and Ferrara (1998) found severity differences in groups of raters across a three-year period that were large enough to affect how students would be classified into performance levels. In their study, the researchers looked at student performance in writing and several other subjects at grades three, five, and eight.

While a number of researchers have conducted studies to examine changes in individual raters' levels of severity over time, the studies report conflicting results. Some researchers contend that the level of severity a rater exercises is a relatively stable effect that changes little over time and is not modifiable by training (Bernardin and Pence, 1980; Lunz and Stahl, 1990; Lunz, Stahl, and Wright, 1996; O'Neill and Lunz, 1996; O'Neill and Lunz, 2000; Raymond, Webb, and Houston, 1991). By contrast, other researchers argue that some raters' levels of severity can shift substantially from reading to reading (Lumley and McNamara, 1993; Myford, Marr, and Linacre, 1996), from essay topic to essay topic (Bridgeman, Morgan, and Wang, 1996; Weigle, 1999), and from day to day within the same reading (Bleistein and Maneckshana, 1995; Braun, 1988; Coffman and Kurfman, 1968; Morgan, 1998; Wilson and Case, 2000; Wood and Wilson, 1974).

Researchers studying the scoring of the free-response sections of AP Examinations have reported differences in the levels of severity that raters exercise. Coffman and Kurfman (1968) noted that the four raters who scored essays written for the AP American History Exam employed different grading standards, some rating more severely than others. Substantial rater severity differences were found in the scoring of five AP Exams, including the AP ELC Exam (Braun, 1988; Braun and Wainer, 1989). Braun also reported the existence of "table effects" in the scoring of the AP ELC Exam (i.e., for some essay questions, the proportion of variance that was due to between-table differences was larger than the proportion of variance due to rater, essay, day, or time of day of the rating). Bridgeman, Morgan, and Wang (1996) reported that rater severity variance was more of a factor in the scoring of essays for the AP ELC Exam than in the scoring of essays included in other AP Exams. However, task difficulty differences were a greater source of score unreliability than rater severity differences in a number of the AP Exams the researchers studied. Myford and Mislevy (1995) used both qualitative and quantitative methods to study rating behavior in the AP Studio Art general portfolio assessment. They conducted interviews with raters about 18 portfolios that received discrepant ratings to gain insights into the kinds of evidence, inference, arguments, and standards that underlie ratings. The results from their FACETS analysis of the rating data revealed differences in the levels of severity that raters exercised in scoring the portfolios.

## *Approaches to Rater Calibration*

Rater calibration methods fall into two categories, based on the timing and underlying assumptions. Training prior to scoring (and continued throughout scoring) is thought to bring raters to consensus in their use of a scoring rubric, thus diminishing differences between them in their interpretation of the rubric. By contrast, statistical methods of correcting for rater effects recognize the established limitations of training procedures to make raters truly interchangeable and the consequent need to adjust scores to take into account differences in rater severity that persist after training.

Traditional training methods begin with the presentation of model papers that both conform to the scoring rubric and those whose "fit" is more problematic (Campbell, 1993). Practice with feedback and instruction on how to interpret the performance data are central (Lunz, Stahl, and Wright, 1996; Rudner, 1992; Wilson and Case, 2000). Discussion that builds a community of like-minded raters is the goal (Campbell, 1993; Pula and Huot, 1993; Wolfe and Kao, 1996). A

unique procedure, recommended by Wolfe and Feltovich (1994), is based on identification of the characteristics of accurate, experienced raters. Novices are then taught the procedural knowledge used by "good" raters in addition to the scoring standards (Wolfe and Kao, 1996). Recommendations have been made to select raters based on the identified characteristics of good raters (Pula and Huot, 1993; Wolfe and Kao, 1996) though it is not clear how these qualities could be discovered in novice raters.

As discussed in the previous section, some researchers contend that even extensive training cannot substantially alter the level of severity a rater exercises. Raymond and Houston (1990) and Lumley and McNamara (1993) see additional purposes for rater training: identifying inconsistent raters, and making raters self-consistent. If raters can be shown to rate consistently, there are a variety of statistical methods of correcting for rater effects that can reduce the impact of score unreliability that is due to systematic rater differences.

Researchers studying AP Examinations have proposed several adjustment procedures. Braun, in his 1988 study of five AP Exams, proposed an adjustment process using "calibration of the levels of two of the different factors contributing to the unreliability: readers and days" (p. 2). Braun found that some raters showed considerable variability in the level of severity they exercised from day to day, which led him to conclude, "these findings suggest that we should explore the possibility of calibrating readers separately each day rather than once overall" (p. 9). Accordingly, the adjustment procedure he employed took into consideration the level of severity each rater exercised each day of a multiday reading. His analysis of readers for the AP English Literature and Composition Exam revealed that "fully one-third (13/36) have average deviations that are 0.5 points or more away from zero" (Braun, 1986, p. 15). The differences between readers for AP American History were even more pronounced: "from one-third to one-half of the readers have deviations at least 0.5 points away from zero" (Braun, 1986, p. 23). Braun reported that by calibrating the raters and adjusting students' scores, the individual gains in score reliability were substantial for AP American History, AP European History, and AP English Literature and Composition (i.e., gains on the order of 20–30 percent). By contrast, the gains in score reliability for AP German and AP Chemistry were negligible.

To implement his calibration procedure, which was based on an ordinary fixed-effects analysis of variance model, Braun (1986) needed to compute various components of variance. However, at the time he carried out his research, the theory of variance component estimation was not well developed for incomplete block designs (i.e., designs in which there is much missing data, since not all raters rate all essays).[1] Braun conducted an experiment during an AP reading, employing a partially balanced incomplete block (PBIB) design that allowed him to calculate unbiased estimates of rater effects, even though each rater did not read all essays. He proposed that during an operational AP reading, such experiments could be carried out using several tables of raters, and the adjustments to students' scores in the operational reading would be made based on the results from these small-scale experiments.

Myford and Mislevy (1995) used a many-faceted Rasch measurement approach to analyze data from the AP Studio Art general portfolio assessment. The FACETS computer program they employed adjusts students' scores for differences in rater severity/leniency. Because seven or more raters contribute to the composite score for each portfolio, differences in levels of rater severity tended to average out, to some extent. However, when the researchers adjusted students' composite scores for severity/leniency differences that remained after this "canceling out" effect had occurred, the adjusted scores of about 1 of every 20 students would have moved them up into the next higher AP grade, and about 1 of 20 would have moved to the next lower grade. No students would have moved more than one AP grade.

Longford (1993) argued that in certain settings, variance due to differences in rater consistency can be much larger than variance due to differences in rater severity and therefore should be taken into consideration in adjusting students' scores. Longford's additive variance components model employs an empirical Bayes framework to estimate variances due to true scores, rater severity, and rater inconsistency. The model then adjusts students' scores using the multiple sources of information it obtains about each rater. Longford (1994a, 1994b) employed his model to study three AP Exams. When he compared the students' operational grades to the grades derived by using the adjusted scores, he found that the percent of students who would have received a different AP grade was 0.13 percent for AP Psychology, 0.12 percent for AP Computer Science, and 0.07 for AP English Language and Composition. Adjustment is particularly important when the free-response section of the exam contributes more to the AP composite score than the multiple-choice section of the exam, Longford noted. It is also more important for

[1] Over the last several years, methodologists have been working to extend generalizability theory to increase its flexibility in terms of its design requirements. See especially Brennan (2001) for a discussion of the most recent developments that allow for a relaxing of the requirements for rating designs.

examinations with lower rater reliability, a characteristic of the AP English Language and Composition essays when compared to the rater reliabilities for the AP Psychology and AP Computer Science essays.

Researchers working with rating data from other settings have devised a variety of regression-based procedures to investigate the rater severity effect and to adjust scores for the impact of this effect. Some have experimented with multivariate analysis procedures for incomplete data to impute ratings (Beale and Little, 1975; Houston, Raymond, and Svec, 1991; Little and Rubin, 1987; Raymond, 1986; Raymond and Houston, 1990). Others have proposed least-squares regression procedures (Cason and Cason, 1984; DeGruijter, 1984; Raymond and Viswesvaran, 1993; Raymond, Webb, and Houston, 1991). In some studies, ordinary least-squares approaches are used (e.g., Lance, LaPointe, and Stewart, 1994), while in other studies weighted least-squares approaches are employed (e.g., Wilson, 1988).

One of the major purposes of this study is to illustrate the use of item response theory for calibrating faculty consultants and adjusting students' scores for faculty consultant severity effects. A recent application of IRT to calibrate raters is reported in Engelhard, Myford, and Cline (2000). This study focused on assessor effects in the National Board for Professional Teaching Standards assessments for Early Childhood/Generalist and Middle Childhood/Generalist. The study provides strong evidence in support of the use of IRT-based methods, such as many-faceted Rasch measurement, for calculating and evaluating the impact of rater severity/leniency.

## *Investigations of Relationships Between Student and Rater Background Characteristics*

Several studies have demonstrated the importance of looking at potential interactions between students and raters. In large-scale writing assessments, female and white students consistently outperform male and minority students (Du and Wright, 1997; Engelhard, 1994; Engelhard, Gordon, and Gabrielson, 1992; Engelhard, Gordon, Siddle-Walker, and Gabrielson, 1994; Gyagenda and Engelhard, 1998). Further, raters evaluating narrative writing samples are remarkably accurate in identifying the gender of the student writers (Gordon and Engelhard, 1995). Peterson and Bainbridge (1999) report that teachers construct writer gender as they read student narratives. While there is no research indicating that student gender, race/ethnicity, or language impact the scoring of high-level academic writing such as the AP ELC Examination, there have been studies that have looked at relationships between student and rater background characteristics in essay scoring.

Wolcott et al. (1988) found strong rater agreement in the scoring of college essays written by native English speakers but discrepant scoring for papers containing English-as-a-second-language (ESL) errors. Though only two raters participated in the study, the researchers indicated that, for high-scoring papers, the presence of ESL errors contributed to discrepancies in scores. In a study of Chinese ESL writers, McDaniel (1985) reported a higher correlation between sentence scores and the number of errors in sentence structure, grammar, and punctuation for the Chinese ESL students than for students whose first language was English. McArthur (1981) found significant interactions between student and rater ethnicity in a study of the scoring of fifth- and sixth-grade essays. Factual narrative essays were read by two Hispanic and two non-Hispanic teacher-raters, who assigned four scores using a six-point rubric. The researcher concluded that essays written by Hispanic students were judged differently when scored by Hispanic and non-Hispanic raters. Similarly, essays written by Hispanic students received different scores, depending upon the cultural background of the rater.

Other writing assessment studies have found little or no relationship between student and rater background characteristics. Chase (1986) gave 83 in-service elementary and middle school teachers a single contrived essay to score and a class record that indicated the gender, race/ethnicity, and level of expectation of the fictitious fifth-grade student who wrote the essay. The student record identified the student as a low or high achiever (i.e., expectation), black or white, male or female. Two versions of the essay were prepared: one using poor handwriting, and the other using good handwriting. Each teacher read the single essay under one combination of these four conditions. The results revealed complex interactions of expectation level, quality of penmanship, and sex within race. However, the mean scores assigned the essay by white and black teachers were not significantly different. Similarly, the mean scores assigned by male and female teachers were not significantly different. Shohamy, Gordon, and Kraemer (1992) also found no effect for rater background in a study of the scoring of twelfth-grade essays written by English-as-a-foreign-language (EFL) students and rated by experienced EFL teachers and non-teachers. Graham and Dwyer (1987) gave two groups of preservice regular education teachers different information about the fourth-grade students who wrote stories. One group was told whether the children who composed the stories were "learning disabled" or "normal." The other group did not receive this information. Additionally, half the teachers attended a brief

training session to learn to use the essay scoring criteria, while the other half of the teachers received more extensive training and practice in using the scoring criteria. The "learning disabled" or "normal" student label appeared to influence the scoring to a small extent, but the group with more extensive training was less influenced by the labeling than the group with minimal training.

Our review of the literature has not identified any relevant studies examining the differential impact of rater severity on essay ratings and the AP grades of student subgroups based on gender, race/ethnicity, and best language.

# Building a Conceptual Model for the Measurement of English Achievement

The conceptual model for the measurement of English achievement used to guide this study is presented in Figure 1. This conceptual model hypothesizes that the student's grade on the AP ELC Exam is the outcome of a set of intervening processes. Ideally, the major determinant of the student's AP grade should be the construct or latent variable being measured: English achievement. However, the student's AP grade is not a direct measure of the construct; rather, the AP grade is, at best, only an indirect measure of the construct. Between the construct to be measured and the actual measure of the construct are a series of intervening processes. If not carried out properly, any one of the intervening processes could introduce construct-irrelevant variance into the assessment. Construct-irrelevant variance is defined as "the degree to which test scores are affected by processes that are extraneous to [the] intended construct" (*Standards for Educational and Psychological Testing*, 1999, p. 10). These sources of unwanted variance, if not monitored carefully, have the potential to threaten the validity of the AP grades. In building a validation argument, a major goal is to determine to what extent AP ELC grades may be "influenced… by components that are not part of the construct" (*Standards for Educational and Psychological Testing*, 1999, p. 10). We have identified some potential sources of construct-irrelevant variance shown in Figure 1. As we present the conceptual model in greater detail in the next section, we will explain how each of these sources could adversely affect the measurement of the construct. We will also point out the steps

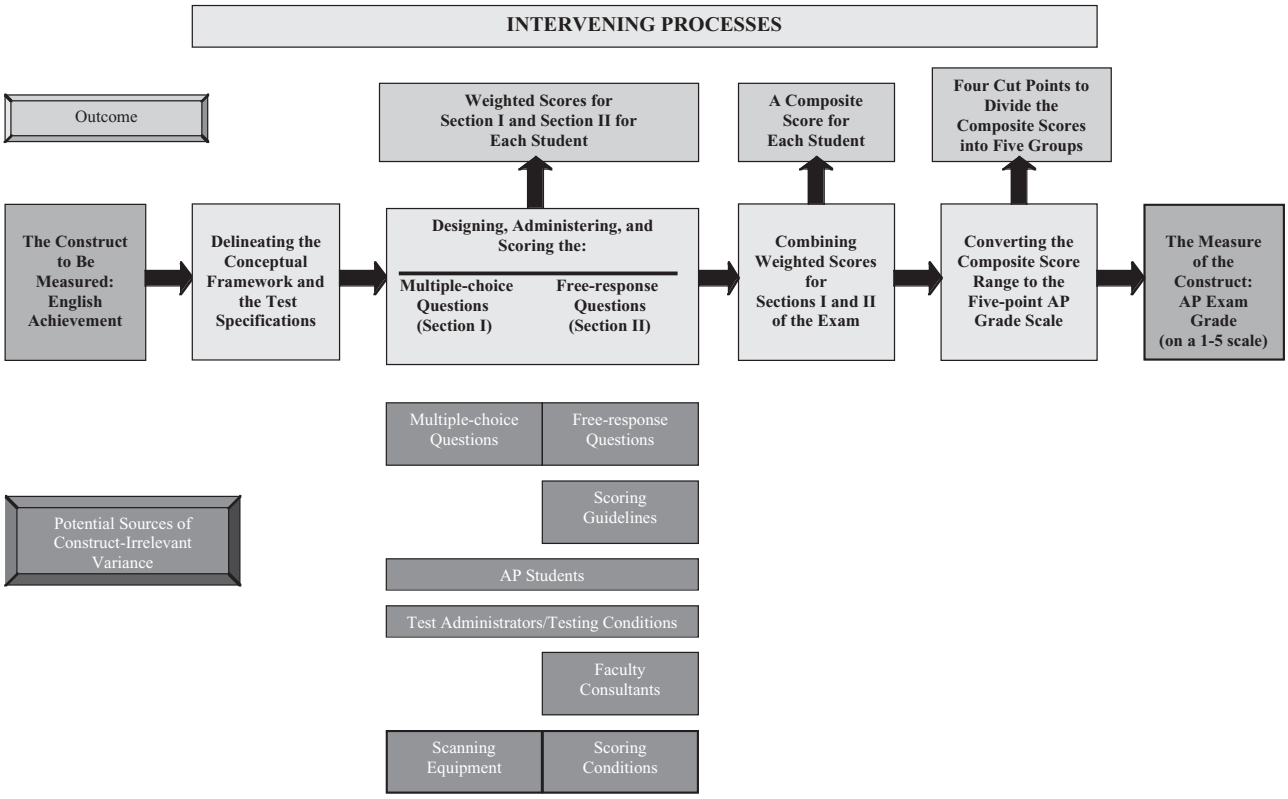


**Figure 1.** A conceptual model for the measurement of English achievement on the AP English Literature and Composition Exam.

the AP Program has taken to help ensure that these sources are not systematically influencing AP grades.

## *Explanation of the Conceptual Model*

For the AP Program, validation begins with a "statement of the proposed interpretation" of AP grades and a "rationale for the relevance of the interpretation to the proposed use" (*Standards for Educational and Psychological Testing*, 1999, p. 9). Students who take AP Exams may be eligible to receive college credit for work they have completed in high school. The College Board (2000) describes the intended use of AP grades:

> Colleges must be reasonably certain that the AP grades they receive represent a level of achievement equivalent to that of students who take the same course in the colleges' own classrooms… the question of greatest concern to colleges–are their AP students who are exempted from introductory courses as well prepared to continue in a subject area as students who took the first course in college? (p. 70)

Establishing the validity of the AP ELC Exam requires examining evidence to determine to what extent the evidence supports this interpretation of AP grades. Accordingly, the process of validation involves gathering and reviewing evidence that supports (or refutes) the proposition that AP students who receive college credit for having taken AP ELC are as well prepared to continue in this content area as students who took an introductory English course in college.

### Defining the Construct, Delineating the Conceptual Framework, and Designing the Exam

The AP English Development Committee, composed of college and university English faculty and high school AP English teachers, works with content experts at ETS to define the scope of the construct; delineate the conceptual framework for the exam; design the test specifications; and develop the test questions, response formats, and scoring guidelines. Colleges and universities that offer an introductory English course are surveyed, and the committee reviews the survey results to establish the content for the AP ELC course. The content domain for the exam is carefully mapped to ensure that the skills and abilities that are to be assessed on the exam are skills and abilities that are developed in students enrolled in introductory college English courses.

The AP ELC Exam consists of two sections. Section I is composed of multiple-choice questions that are designed to assess "the student's critical reading of selected passages" (The College Board, 2000, p. 45). Section II contains free-response questions that function "as a direct measure of the student's ability to read and interpret literature and to use other forms of discourse effectively" (The College Board, 2000, p. 45). The multiple-choice questions, free-response questions, response formats, and the scoring guidelines used to evaluate students' written responses provide the operational definition of the English achievement construct.

*The Standards for Educational and Psychological Testing* (1999) describe several steps in the assessment design process that, if not carried out appropriately, could compromise validity: (1) the construct to be measured may be inadequately conceptualized, (2) the conceptual framework may inadequately specify the aspects (e.g., content, skills, processes, and diagnostic features) of the construct, (3) the construct may be inadequately represented by the questions included on the exam, (4) the exam may fail to capture critical aspects of the construct (i.e., construct underrepresentation), (5) certain questions appearing on the exam may not meet the requirements set forth in the test specifications, and (6) certain scoring criteria may be inconsistent with the exam's purposes. In building a validity argument for the AP ELC Exam, evidence has been gathered to demonstrate that the assessment design process is carried out in a credible, defensible fashion that meets the standards set forth in the *Standards for Educational and Psychological Testing* (1999). The test developers have documented the test specifications, including a rationale and a description of the process used to develop them (Standard 3.3). Content experts who are not ETS employees have reviewed the test specifications, and the results of that review have been documented (Standard 3.5). The test developers have documented the extent to which the content domain represents the English achievement construct and the test specifications (Standard 3.11). With the development of each test form, a diverse group of external content experts carefully reviews the test questions, response formats, and scoring guidelines to ensure that they adequately represent the defined content domain, are not technically flawed, and meet the content and statistical requirements set forth by the test specifications (Standards 3.6 and 3.11). They classify the questions included on each test form, using the categories included in the test specifications (Standard 3.7), and check to make certain that the questions and scoring criteria are consistent with the exam's purpose (Standards 3.6 and 3.14). (See "How AP Courses and Exams are Developed" from the *AP Technical Manual* at http://www.collegeboard.com/ap/techman/chap2/ for a description of the exam development process. See also Chapter 2, "Validity of Advanced Placement Grades," of the *College and*

*University Guide to the Advanced Placement Program* for an explanation of the process used to establish the content validity of AP Examinations.)

Certain **questions** or **scoring guidelines** may contain construct-irrelevant aspects related to content, response format, or scoring criteria that might differentially affect the scores of one or more subgroups of test-takers, confounding the construct definition (Standard 3.6). The AP test development process includes a fairness review of all multiple-choice and free-response questions, response formats, and scoring guidelines prior to the operational administration of the test. A diverse group of external content experts and ETS test developers who have received special training in fairness review procedures scrutinizes the test materials for language that might be interpreted differently by members of different subgroups. They also look for material that may be inappropriate, confusing, or potentially offensive to one or more subgroups of AP test-takers. Additionally, AP statisticians conduct DIF analyses on operational AP Exams to screen for improperly functioning multiple-choice questions. If after the administration of the exam there are multiple-choice questions that are found to show DIF across gender, racial/ethnic, or linguistic groups and the external content experts can attribute the differences in performance to factors other than the knowledge that is being tested, then those questions are eliminated from scoring. (See "Differential Item Functioning" from the *AP Technical Manual* at http://www.collegeboard.com/ap/techman/chap4/differential.htm for an explanation of how the AP Program conducts DIF analyses and uses the results from those analyses.)

## Administering the Exam

Students take the three-hour AP ELC Examination in May each year. In order to avoid introducing construct-irrelevant variance into the assessment process, AP students need to clearly understand the purpose of the exam, the testing process, how the test will be administered (i.e., choice of test formats), the time limits, and the testing instructions. They need to be given information about the advisability of omitting responses (Standards 8.1, 8.2, and 8.3). Additionally, where appropriate, students should be provided in advance with information about the content of the test, including topics covered, item formats, and test scoring criteria (Standard 8.2). Students can access information about the AP ELC course and exam at the College Board Web site (http://apcentral.collegeboard.com/article/0,1281,151-162-0-4499,00.html). The course description can be downloaded free of charge, as can information explaining the two sections of the exam (i.e., the purpose of each section, the time limits, how much each section contributes to the total grade). Sample multiple-choice and free-response questions from the previous year's exam are posted at the Web site. Additionally, the scoring guidelines used to evaluate students' responses to the free-response questions are provided, as well as samples of student essays written for each question. The Web site also contains useful information about study and test-taking strategies for AP ELC. Students enrolled in the AP ELC course are encouraged to take part in an online moderated discussion group maintained by the College Board. Several resources are available to students for a fee. Students can purchase the English APCD®, an exam review tool that provides guidance and practice to prepare for the AP Exam. They can practice taking essays under standardized test conditions and submit their essays to the College Board online evaluation service to receive detailed feedback about their performance. Each year the AP Program makes available for purchase the previous year's exam. The released exam includes all the test questions, the scoring guidelines used to evaluate the students' essays, the answer key for the multiple-choice questions, sample student responses to the essay questions, and commentaries on the students' essays. Finally, a number of companies (e.g., Princeton Review, CliffsNotes, Barron's) market test preparation materials to help students get ready to take the AP Exams. These publications contain practice exams and provide advice on test-taking strategies.

**Test administrators** can also introduce construct-irrelevant variance into the assessment process if they have not been properly trained to administer the assessment, or if they do not carefully follow standardized administration procedures (Standards 5.1 and 13.10). Disruptive **testing conditions**, such as distracting noises, a room that is too hot or too cold, or a poorly lit room, can also introduce construct-irrelevant variance (Standard 5.4). The AP Program strives to ensure that those persons who administer AP Exams in the schools are proficient in exam administration procedures and that they understand the importance of strictly following the instructions for exam administration that the AP Program provides.

## Scoring the Exam

In June, college and university English professors who teach introductory college English courses and experienced high school AP English teachers assemble to score students' essays for Section II of the exam. These faculty consultants carefully study the scoring guidelines and then receive intensive training to become familiar with the demands of the particular free-response question

they are to score. As part of their training, they learn to apply the scoring guidelines to score students' essays. The faculty consultants practice scoring preselected essays until they develop a shared understanding of how to apply and interpret the scoring guidelines. Since the scoring of AP Exams involves human judgment, it is important that those scoring the exam strictly adhere to the scoring criteria they have been given (Standard 5.9). If there were faculty consultants who after training rated significantly more leniently (or, conversely, more harshly) than other faculty consultants, that could introduce construct-irrelevant variance into the ratings, particularly if there were faculty consultants that showed systematic patterns of differential severity when rating male and female students, various racial/ethnic groups, or different linguistic groups. In this situation, how well a student performed on Section II of the exam could be heavily dependent upon the particular set of faculty consultants who, by luck of the draw, happened to rate the student's essays. Similarly, if there were faculty consultants who, even after intensive training, were unable to use the scoring guidelines consistently when evaluating students' essays, or faculty consultants who allowed personal biases to cloud their judgments, they would introduce additional construct-irrelevant variance into the assessment process. The quality of the ratings they provide would be highly suspect. The AP Program has carefully documented the processes employed to select, train, and qualify the faculty consultants (Standard 3.23) and each year publishes, as part of the released exam, samples of the training materials the faculty consultants used to learn to apply the scoring guidelines (Standard 3.24). Additionally, the AP Program has instituted a series of checks and balances to help ensure that the essay grading is carried out in a fair and consistent manner. (See "Scoring the Free-Response Section" from the AP Technical Manual at http://www.collegeboard.com/ap/ techman/chap3/score-fc.htm for a description of the quality control monitoring steps that are followed at each reading.)

**Scoring conditions** can also introduce construct-irrelevant variance. The faculty consultants score students' essays for a week, reading eight hours a day (with breaks). Distracting noises, a room that is too hot or too cold or poorly lit might differentially affect faculty consultant performance. Fatigue and/or boredom may also differentially affect their performance. Some faculty consultants may be prone to fatigue and/or boredom, and, as a reading progresses, the quality of their ratings may suffer. While they may have used the scoring guidelines appropriately early on in the reading, as they tire, their attention may wane, and they may lose focus and begin to use the scoring guidelines inappropriately (or, perhaps, inconsistently) across essays. Other faculty consultants may not be as affected by either fatigue or boredom and thus may be able to maintain the quality of their ratings, continuing to employ the scoring guidelines appropriately for extended periods of time. Table leaders who supervise faculty consultants during an AP reading receive summary statistical information twice a day concerning the scores that each faculty consultant has given. Having access to this information makes it possible for the table leaders to provide real-time feedback to those faculty consultants who may be experiencing adverse effects from fatigue or boredom. Table leaders also engage in back reading (i.e., they reread selected essays that each of the faculty consultants at their table has previously scored) as a check on accuracy. Periodically, the AP Program conducts special studies to investigate the impact of time of day and day of week on the reliability of the scoring. (See "Scoring the Free-Response Section" and "Scoring Reliability Studies" from the *AP Technical Manual* at http://www.collegeboard.com/ap/techman/chap3/score-fc.htm for a description of the methods used to attempt to ensure that all faculty consultants are using the scoring guidelines appropriately.)

The multiple-choice questions included on the AP ELC Exam are scored by machine. If the **scanning equipment** were to malfunction, that could introduce construct-irrelevant variance into the assessment process. Each answer sheet is fed through an electronic scanner, creating a record for the student by transferring information directly to cartridges. The computer processes the scanning cartridge, checking for invalid and missing data, and scores the students' responses to the multiple-choice items. (See "Scoring the Multiple-Choice Section" from the *AP Technical Manual* at http://www.collegeboard.com/ap/techman/chap3/score mc.htm for a description of the scanning process.) Furthermore, the Reader Management System, the scoring service for AP Exams, has documented the procedures the system uses to detect inaccurate scoring and has detailed the steps that are to be followed when there are found to be machine scoring errors (Standard 5.8).

### Combining Weighted Section Scores to Produce Composite Scores

Students taking Section I of the AP ELC Exam respond to 55 multiple-choice questions. When students take Section II of the exam, they write essays for three free-response questions. Performance on the two sections of the exam is not equally weighted. A computer mechanically carries out the computation of composite scores after the faculty consultants have scored the free-response section of the exam. (For a detailed description

of this statistical process, see Released Exam—1999 AP English Literature and Composition, The College Board, 1999a, pp. 71–72.)

### Converting Composite Scores to AP® Grades

After the composite scores have been calculated, the scores are then converted to the five-point AP grade scale. The conversion process involves establishing four cut points, dividing the composite scores into five groups. Those who set the cut points make use of a number of pertinent information sources to help ensure continuity of AP grades across years. Each student receives a single score on the 1–5 AP grade scale. The scale transformation and setting of grade boundary ranges depend heavily on the equating of multiple-choice scores on the previous year's exam to the multiple-choice scores on the present form through a set of link items that are common to both test forms. (See "Calculating the AP Grade" from the AP Technical Manual at http://www.collegeboard.com/ap/techman/chap3/grade.htm for a description of the grade setting process, and "How AP Grades Are Determined" from the Released Exam—1999 AP English Literature and Composition, The College Board, 1999a, pp. 71–72.)

## Building a Psychometric Model from the Conceptual Model

In this study, we employed a psychometric model that enabled us to operationalize key components of the conceptual model. Our psychometric model provides practical, useful information about some (but not all) of the intervening processes. By reviewing output from our analyses, we are able to monitor the extent to which these intervening processes may be introducing construct-irrelevant sources of variance into the assessment.

Our psychometric model includes *questions* as one facet in our analyses. Each separate multiple-choice and free-response question is an "element" of the questions facet. From our analyses, we obtain a measure of the difficulty of each question (i.e., for multiple-choice questions, how hard it is for the student to get the answer correct; for free-response questions, how hard it is for a student to receive a high rating on the question) and a standard error for the measure. We also obtain an estimate of the extent to which the question "fits" with the other questions included on the test—that is, whether the questions work together to define a single unidimensional construct, or whether there is evidence that the questions are measuring multiple, independent dimensions of a construct (or, perhaps, multiple constructs). The fit statistics for an individual question signal the degree of correspondence between the students' performances on that particular question when compared to their performances on the other questions. They provide an assessment of whether scores on the questions can be meaningfully combined to produce a single composite score, or whether there may be a need for separate scores to be reported rather than a single composite score.

A second facet included in our psychometric model is *faculty consultants*. From our analyses, we obtain a measure of the level of severity each faculty consultant exercised when evaluating students' essays and a standard error of that measure. This information can be used to judge the degree to which faculty consultants are functioning interchangeably. The computer program also provides an estimate of the consistency with which a faculty consultant applies the scoring guidelines. Faculty consultant "fit" statistics are estimates of the degree to which a faculty consultant is internally consistent when using the scoring guidelines to evaluate multiple essays.

A third facet in our psychometric model is *students*. The output from our analyses provides a measure of each student's level of English achievement on the AP ELC Exam and a standard error of the measure. In producing these estimates, the computer program adjusts the student's AP grade for the level of severity that the particular faculty consultants scoring that student's three essays exercised. In effect, the program "washes out" these unwanted sources of construct-irrelevant variance. The resulting AP grade reflects what the student would have received if faculty consultants of average severity had rated the student's essays. The computer program also produces student "fit" statistics that are indices of the degree of consistency shown in the evaluation of the student's level of English achievement across questions and across faculty consultants. Through fit analyses, students who exhibit unusual profiles of scores across questions (i.e., students who appear to do unexpectedly well [or poorly] on some questions) can be identified. Flagging the AP grades of these misfitting students allows for an important quality control check before grade reports are issued—an independent review of each misfitting student's performance across questions. One can determine whether the particular question-level scores that the computer program has identified as surprising or unexpected for that student are perhaps due to random (or systematic) error and then use that information to decide whether those question-level scores might

need to be changed. Alternatively, through fit analysis, one might determine that, indeed, the student performed differentially across questions, and the question-level scores should be left to stand as they are.

The psychometric model also takes into consideration the structure of the scoring guidelines used to evaluate students' essays. These analyses provide useful information that enables the determination of whether or not the scoring guidelines are functioning as intended. For example, by examining the rating scale category thresholds that the computer program produces, it is possible to determine whether the nine categories in the AP scoring guidelines for a free-response question are appropriately ordered and are clearly distinguishable.

Finally, through bias analyses, the psychometric model determines whether faculty consultants are differentially severe when evaluating the essays of different subgroups of students. From our analyses, we are able to determine whether there are any faculty consultants that show systematic patterns of differential severity or leniency when rating males and females, various racial/ethnic groups, or different linguistic groups (i.e., systematic interaction effects between faculty consultants and various subgroups of test-takers).

Our psychometric model provides useful information about the test questions, the faculty consultants, the students, the scoring guidelines, and various interactions between facets in the model (i.e., faculty consultants and students). However, the psychometric model does not provide any information about the adequacy of the initial conception of the construct (i.e., whether the construct is appropriately defined and bounded; how well the content domain was mapped; whether the skills and abilities assessed on the exam are skills and abilities that are developed in students enrolled in introductory college English courses; whether the test specifications adequately delineate the question format, the response format, and the scoring procedures; whether the questions included on the exam adequately represent the construct; whether the questions meet the requirements of the test specifications; whether the scoring guidelines are in sync with the purpose of the exam, etc.). The psychometric model also provides no direct information about the appropriateness of the algorithm for combining weighted scores or the process of converting the composite score range to the AP grade scale. Our analyses are not useful for monitoring these intervening processes. Other approaches are needed to gather evidence to rule out the possibility that these processes are introducing construct-irrelevant sources of variance into the assessment, threatening the validity of the AP grades.

# A Many-Faceted Rasch Measurement Approach to the Measurement of English Achievement

The procedures described in this section for examining the quality of the ratings assigned by AP ELC faculty consultants are based on a many-faceted version of the Rasch measurement model for ordered response categories developed by Linacre (1989) and implemented in the FACETS computer program. Many-faceted Rasch measurement models are extended versions of the basic one-parameter Rasch model (Andrich, 1988; Rasch, 1980; Wright and Masters, 1982). Researchers using a many-faceted Rasch measurement (MFRM) approach establish a statistical framework for analyzing their rating data, summarizing overall rating patterns in terms of group-level main effects for variables (or "facets") of their rating operation, such as faculty consultants, students, and questions. When a MFRM analysis is run, the contribution of each facet can be separated out and examined independently of other facets to determine to what extent the various facets are functioning as intended. Using an MFRM approach also allows researchers to look at individual-level effects of the various "elements" within a facet (that is, how individual faculty consultants, students, or questions included in the analysis are performing). The ability to obtain valuable individual-level diagnostic information about how each particular element within the rating operation is functioning sets a MFRM approach apart from other ANOVA-based or regression approaches to analyzing rating data (e.g., generalizability analyses, ordinary least-squares regression, weighted least-squares regression, etc.). (For a discussion of the differences between the MFRM approach and other ANOVA-based approaches to analyzing rating data, see Wilson and Case, 2000, pp. 116-117.)

When a MFRM analysis is run, the various facets are analyzed simultaneously but statistically independently and are calibrated onto a single linear scale (i.e., the logit scale). The joint calibration of facets makes it possible to measure faculty consultant severity on the same scale as student English achievement and question difficulty. All facets of the rating operation are expressed in a common equal-interval metric (i.e., log-odds units, or logits). A MFRM model is essentially an additive linear model that is based on a logistic transformation of observed ratings to a logit or log-odds scale. The logistic transformation of ratios of successive category probabilities (log odds) can be viewed as the dependent variable with various facets, such as students, questions, and faculty

consultants, conceptualized as independent variables that influence these log odds. If the rating data show sufficient fit to the model, then researchers are able to draw useful, diagnostically informative comparisons among the various facets, which is another way in which a MFRM approach differs from other ANOVA-based approaches to analyzing rating data. Results from generalizability and least-squares analyses are expressed in the original raw score units—a nonlinear ordinal scale metric.[2]

In this study, the many-faceted Rasch measurement model takes the following form:

$$ln[P_{nijk} / P_{nijk-1}] = \Theta_n - \xi_i - \alpha_j - \tau_k, \quad (1)$$

where

$P_{nijk}$ = probability of student *n* receiving a rating of *k* on question *i* from faculty consultant *j*,

$P_{nijk-1}$ = probability of student *n* receiving a rating of *k – 1* on question *i* from faculty consultant *j*,

$\Theta_n$ = English achievement for student *n*,

$\xi_i$ = difficulty of question *i* (including multiple-choice and free-response questions),

$\alpha_j$ = severity of faculty consultant *j*, and

$\tau_k$ = difficulty of receiving a rating of *k* relative to a rating of *k – 1*.

The category coefficient $\tau_k$ is not considered a facet in the model. The function of this parameter is to indicate to FACETS how the rating data are to be handled. In this case, the parameter specifies that a rating scale model (Andrich, 1978) should be used; that is, in the analysis, FACETS was directed to treat the nine-category scoring guidelines for the three free-response questions as if all shared the same rating scale structure, with category coefficients calibrated jointly across the three free-response questions. (The category coefficient is understood to be zero for the multiple-choice questions.) For the three free-response questions, a category coefficient reflects the difficulty of moving across adjacent categories of the scoring guidelines.

Based on the FACETS model presented in Equation 1, the probability of student *n* with level of English achievement $\Theta_n$ obtaining a score of *k* (*k* = 1, … , m) on question $\xi_i$ from faculty consultant $\alpha_j$ with a category coefficient of $\tau_\kappa$ is given as

$$P_{nijk} = exp\ [k\ (\Theta_n - \xi_i - \alpha_j) - \sum_{h=1}^{k} \tau_h\ ] / \gamma, \quad (2)$$

where $\tau_1$ is defined to be 0, and $\gamma$ is a normalizing factor based on the sum of the numerators.

For each element of each facet, a MFRM analysis provides a measure (a logit estimate of the calibration), a standard error (information about the precision of that logit estimate), and fit indices (information about how well the data fit the expectations of the measurement model). Fit indices indicate the degree to which observed ratings match the expected ratings that are generated by the MFRM model. Large differences between the observed and expected ratings (expressed as standardized residuals) indicate surprising or unexpected results. Useful indices of psychometric quality can be obtained by a detailed examination of the standardized residuals calculated as

$$Z_{nij} = (x_{nij} - E_{nij}) / [\sum_{k=1}^{m} (k - E_{nij})^2\ P_{nijk}\ ]^{1/2}, \quad (3)$$

where

$$E_{nij} = \sum_{k=1}^{m} k\ P_{nijk} \quad (4)$$

The standardized residuals $Z_{nij}$ can be summarized over different facets and different elements within a facet in order to provide indices of model–data fit. These residuals are typically summarized as mean-square error statistics called OUTFIT and INFIT statistics. The OUTFIT statistics are unweighted mean-

[2] Raw scores are not expressed on a linear scale. Rather, raw scores are distorted at both ends of the variable, resulting in an inability to maintain a constant unit of measurement along the entire continuum. In this situation, a difference of one score point on the left end of the student proficiency continuum does not have the same meaning as a difference of one score point on the right end of that continuum, making differences between students at the extremes look smaller than they would look if the test items were centered on the extreme students. This distortion of the score scale can thus lead to an inconsistency in the ordering of students by their proficiencies. Raw scores need to be re-expressed on a linear scale if one wishes to carry out arithmetic operations on those scores, since even the calculation of simple statistics such as means and standard deviations are based on the assumption of linearity. To remedy this, the raw scores can be transformed to a "logit" metric. This straightening process, in effect, gets rid of the distortions at the extremes of the continuum, making it possible to display the variable in a form that is not dependent upon how the test items are targeted on students (Wright and Masters, 1982, pp. 31–34). While raw score nonlinearity may not be obvious in a given set of test items or student scores, the nonlinearity does become obvious when one obtains item scores from samples of students having different levels of proficiency, or when one obtains student scores from subsets of test items having different levels of difficulty. As Wright and Stone (1979) sum it up, "Although test scores usually estimate the order of persons' abilities rather well, they never estimate the spacing satisfactorily. Test scores are not linear in the measures they imply and for which they are used" (p. 7).

squared residual statistics that are particularly sensitive to outlying unexpected ratings. The INFIT statistics are based on weighted mean-squared residual statistics and are less sensitive to outlying unexpected ratings. Engelhard (in press; 1994) and Myford and Wolfe (2001) provide a description regarding the interpretation of these fit statistics within the context of rater-mediated assessments.

To be useful, an assessment must be able to separate students by their performance (Stone and Wright, 1988). FACETS produces a *student separation ratio*, $G_N$, which is a measure of the spread of the student English achievement measures relative to their precision. Separation is expressed as a ratio of the "true" standard deviation of student English achievement measures (i.e., the standard deviation adjusted for measurement error) over the average student standard error (Equation 5):

$$G = \text{True SD} / \text{RMSE} \quad (5)$$

where True SD is the standard deviation of the student English achievement measures corrected for measurement error inflation, and RMSE is the root mean-square error, or the "average" measurement error of the student English achievement measures.

Using the student separation ratio, one can then calculate the student separation index, which is the number of measurably different levels of student performance in the sample of students. For example, a student separation index of 3 would suggest that the assessment process is sensitive enough to be used to separate students into three statistically distinct groups. According to Fisher (1992), the functional range of person measures is generally around four true standard deviation units. When computing the student separation index, Fisher suggested that the functional range should be inflated by one RMSE to allow for error in the observed measures. If one defines statistically distinct levels of student achievement as achievement strata with centers that are three measurement errors apart (Wright and Masters, 1982), then the student separation index can be computed using Equation 6:

$$(4 \text{ True SD} + \text{RMSE}) / (3 \text{ RMSE}) = (4G + 1) / 3. \quad (6)$$

Similarly, FACETS produces a *faculty consultant separation ratio*, which is a measure of the spread of the faculty consultant severity measures relative to their precision. The *faculty consultant separation index*, derived from that separation ratio, connotes the number of statistically distinct levels of severity in the sample of faculty consultants. A separation index of 1 would suggest that all faculty consultants were exercising a similar level of severity and could be considered as one interchangeable group. We will be reporting student and faculty consultant separation indices in our results, but not their associated separation ratios. The separation indices are more readily understood and have more practical utility, in our view.

Another useful statistic is the *reliability of separation index*. This index provides information about how well the elements within a facet are separated in order to define reliably the facet (e.g., the students, the faculty consultants). This index is analogous to traditional indices of reliability, such as Cronbach's coefficient alpha and KR20, in the sense that it reflects an estimate of the ratio of "true" score to observed score variance, where

$$(\text{True SD})^2 = (\text{Observed SD})^2 - (\text{RMSE})^2. \quad (7)$$

In equation 7, *Observed SD* is the observed standard deviation of the student achievement measures (or, in the case of faculty consultant reliability, the observed standard deviation of the rater severity measures). Separation reliability can then be calculated as

$$R = (\text{True SD})^2 / (\text{Observed SD})^2 = G^2 / (1 + G^2). \quad (8)$$

When G = 1, then the True SD = RMSE, and the resulting separation reliability is 0.5. If separation reliability is less than 0.5, then that denotes that the differences between the measures are due mainly to measurement error (Fisher, 1992). Andrich (1982) provides a derivation of this reliability of separation index. Detailed general descriptions of the separation statistics are also provided in Wright and Masters (1982) and Fisher (1992). The reliability of separation indices have slightly different substantive interpretations for different facets in the model. For students, the reliability of separation index is comparable to coefficient alpha, indicating the reliability with which the assessment separates the sample of students (that is, the proportion of observed sample variance which is attributable to individual differences between students, Wright and Masters, 1982). Unlike interrater reliability, which is a measure of how *similar* rater severity measures are, the student separation reliability is a measure of how *differen*t the student English achievement measures are (Linacre, 1998). By contrast, for faculty consultants, the reliability of separation index reflects potentially unwanted variability in severity.

Equation 2 can be used to generate a variety of expected scores under different conditions reflecting various assumptions regarding the assessment process. For example, it is possible to estimate an expected rating for a student on a particular free-response question that would be

obtained from a faculty consultant who exercised a level of severity equal to zero (i.e., a faculty consultant who was neither more lenient nor more severe than other faculty consultants). In this case, the faculty consultant, *j*, would be defined as $\alpha j = 0$, and the adjusted probability, *AP*, and adjusted rating, *AR*, would be calculated as follows:

$$AP_{nijk} = exp\ [k(\Theta_n - \xi_i - 0_j) - \sum_{k=1}^{m} \tau_k] \ / \ \gamma, \quad (9)$$

and

$$AR_{nij} = \sum_{k=1}^{m} k\ AP_{nijk} \quad (10)$$

Equation 10 can be interpreted as producing an expected adjusted rating for student *n* on free-response question *i* from faculty consultant *j* ($\alpha_j = 0$).

To investigate whether faculty consultants were differentially severe when evaluating the essays of different subgroups of students, three additional facets were added to the model in Equation [1]. Specifically, facets were added to represent student gender, student race/ethnicity, and student first language so that bias interaction analyses could be performed.

# Method

## *Participants*

### Students

The total pool for the 1999 AP ELC Exam was 174,857 students. The background characteristics for these students are shown in Table 1. Sixty-four percent were female, and 36 percent were male. Seventy-four percent of the students were white, 10 percent were Asian/Asian American/Pacific Islander, 5 percent were Black/African American, 4 percent were Mexican American/Chicano, 3 percent were South American/ Latin American/Central American/Other Hispanic, 1 percent were Puerto Rican, 1 percent were American Indian/Alaskan Native, and 4 percent identified themselves as Other. For 94 percent of the students, English is their first and "best" language. Six percent consider themselves equally proficient in English and one other language, and 1 percent speak a language other than English as their first language. For comparison purposes, Table 1 also includes the background characteristics for the 5 percent sample used in this study. As the table shows, the total pool is well represented by the 5 percent sample.

TABLE 1

**Background Characteristics of Students Taking the 1999 AP English Literature and Composition Exam**

| | *Total Population* | | *Sample (5 percent)* | |
|---|---|---|---|---|
| | *Freq.* | *Percent* | *Freq.* | *Percent* |
| ***Gender*** | | | | |
| Female | 111,186 | 63.6 | 5,574 | 64.5 |
| Male | 63,669 | 36.4 | 3,068 | 35.5 |
| ***Race/Ethnicity*** | | | | |
| 1 American Indian or Alaska Native | 933 | 0.6 | 42 | 0.5 |
| 2 Black or African American | 8,812 | 5.4 | 410 | 5.1 |
| 3 Mexican American or Chicano | 5,756 | 3.5 | 266 | 3.3 |
| 4 Asian, Asian American, or Pacific Islander | 16,384 | 10.0 | 840 | 10.4 |
| 5 Puerto Rican | 1,090 | 0.7 | 55 | 0.7 |
| 6 South American, Latin American, Central American, or Other Hispanic | 4,194 | 2.6 | 194 | 2.4 |
| 7 White | 120,486 | 73.7 | 5,997 | 74.3 |
| 8 Other | 5,915 | 3.6 | 266 | 3.3 |
| ***Best Language*** | | | | |
| 1 English | 159,281 | 93.6 | 7,894 | 93.7 |
| 2 English and another language about the same | 9,840 | 5.8 | 491 | 5.8 |
| 3 Another language | 1,022 | 0.6 | 36 | 0.4 |
| ***AP Grade*** | | | | |
| 1 No recommendation | 9,369 | 5.4 | 377 | 4.4 |
| 2 Possibly qualified | 46,193 | 26.4 | 2,262 | 26.2 |
| 3 Qualified | 61,815 | 35.4 | 3,028 | 35.0 |
| 4 Well Qualified | 37,686 | 21.6 | 1,933 | 22.4 |
| 5 Extremely well qualified | 19,791 | 11.3 | 1,042 | 12.1 |
| **Total** | 174,857 | | 8,642 | |

### Faculty Consultants

There were 612 AP ELC faculty consultants who scored the 1999 AP ELC Exam. When we drew the 5 percent random student sample we used in this study, it included 605 of these faculty consultants. Table 2 compares the background characteristics for the total pool of 612 faculty consultants and for the 605 faculty consultants included in our sample. The background characteristics of the faculty consultants included in the 5 percent sample closely match those of the total pool. In the total faculty consultant pool, 38 percent were male, and 62 percent were female. Forty-nine percent were college instructors of English, and 51 percent were high school AP English teachers. Seventy-eight percent were experienced readers (i.e., had participated in at least one previous AP ELC reading), while 22 percent were first time readers. Seventy-nine percent were white, 5 percent were Black/African American, 2 percent were Asian/Asian American/Pacific Islander, and less than 1 percent were American Indian/Alaskan Native, Mexican American/Chicano, or Latin American/ South American/Central American/Other Hispanic. Eleven percent were of unidentified race/ethnicity.

## *AP English Literature and Composition Examination*

The 1999 AP ELC Exam consisted of two sections. Section I contained 55 multiple-choice questions. Performance on this section accounted for 45 percent of a student's total exam score. The College Board (1999a) describes the content and format of Section I of the 1999 exam:

> Section I requires students to read carefully four to six texts (poems or prose passages) and to answer multiple-choice questions about their content, structure, and style. Although the questions test a student's ability to construe meaning, they also require the candidate to respond to stylistic and structural features of the text (such as patterns of imagery, the use of contrast and repetition), to understand figurative language, and to identify rhetorical or poetic devices. (p. 7)

Section II consisted of three free-response questions. Performance on Section II accounted for 55 percent of a student's total exam score.[3] The College Board (1999a) describes the purpose of the free-response questions:

TABLE 2

**Background Characteristics of Faculty Consultants Scoring the 1999 AP English Literature and Composition Exam**

| | *Total Population* | | *Sample (5 percent)* | |
|---|---|---|---|---|
| | *Freq.* | *Percent* | *Freq.* | *Percent* |
| ***Gender*** | | | | |
| Male | 231 | 37.74 | 227 | 37.52 |
| Female | 381 | 62.25 | 378 | 62.48 |
| ***Race/Ethnicity*** | | | | |
| American Indian or Alaska Native | 1 | 0.16 | 1 | 0.16 |
| Black or African American | 29 | 4.74 | 29 | 4.79 |
| Asian, Asian American, or Pacific Islander | 13 | 2.12 | 13 | 2.15 |
| Mexican American or Chicano | 4 | 0.65 | 4 | 0.66 |
| Puerto Rican | 0 | 0 | 0 | 0 |
| Latin American, South American, Central American, or Other Hispanic | 4 | 0.65 | 4 | 0.66 |
| White | 485 | 79.25 | 479 | 79.17 |
| Other | 8 | 1.31 | 8 | 1.32 |
| (Missing) | (68) | (11.11) | (67) | (11.07) |
| ***Institution*** | | | | |
| College instructor of English | 301 | 49.18 | 296 | 48.93 |
| High school AP English teacher | 311 | 50.82 | 309 | 51.07 |
| ***Reader Status*** | | | | |
| Experienced reader | 476 | 77.78 | 469 | 77.52 |
| New reader | 136 | 22.22 | 136 | 22.48 |
| **Total** | 612 | 100.00 | 605 | 100.00 |

[3] For the 1999 AP English Literature and Composition Exam, the composite score weighting was 1.2272 (I) + 3.05556 [(EPT1) + (EPT2) + (EPT3)], where I is the score on Section I of the exam, EPT1 is the first free-response question, EPT2 is the second free-response question, and EPT3 is the third free-response question (The College Board, 1999b). Using this weighting scheme, Section I of the exam contributed 45 percent toward the composite score, while Section II contributed 55 percent toward the composite score. The maximum possible composite score was 150.

> On the Literature exam, students are typically required to write analytical essays on both a poem and a prose text and to apply a critical generalization about literature to a specific, appropriate text of their own choosing. The essay format allows them to demonstrate skills of organization, logic, and argument to produce a personal discussion of the text. They are also free to select aspects of the passage or poem relevant to their argument and to support their point of view with pertinent evidence.
>
> Essays also allow students to demonstrate their writing skills, which include control of syntax and grammar and breadth and exactness of vocabulary as well as the elements of composition mentioned above. Essays provide students with an opportunity to make their individual voices heard and to show the extent to which they have come to employ a mature and effective style. (p. 2)

## *AP English Literature and Composition Examination Process*

Students took the three-hour 1999 AP ELC Exam in May under standardized test conditions in high schools and other designated testing centers. The exam was administered and proctored by trained AP test administrators. Students were given an hour to complete Section I of the exam (the multiple-choice questions) and two hours to complete Section II (the free-response questions). Students marked their responses to the multiple-choice questions on answer sheets and wrote their essays in a test booklet. (The three free-response questions are included in the Appendix as Tables A1, A3, and A5.)

## *AP English Literature and Composition Scoring Process*

Students' answer sheets were run through an electronic scanner and then scored by computer. The computer calculated a total multiple-choice score for each student by counting the number of questions answered correctly, the number answered incorrectly, and then deducting a fraction of the incorrectly answered questions from those answered correctly to eliminate any benefit gained from random guessing.

The scoring of the students' responses to the three essay questions on the examination took place in Daytona Beach, Florida, in June 1999. Faculty consultants were nested within question; that is, each faculty consultant read essays for only one of the three free-response questions. No faculty consultant read essays for more than one free-response question.

The AP ELC Development Committee prepared a first draft of the scoring guidelines for each of the three free-response questions at the time that the questions were created. After the exam was given, the scoring guidelines were reviewed and refined and then tested by evaluating a randomly selected set of student essays. (See Appendix, Tables A2, A4, and A6, for copies of the scoring guidelines for the three free-response questions.)

During the first morning of the reading, the faculty consultants participated in a training program to prepare them to score students' essays. They reviewed the scoring guidelines and read sample benchmark papers that had previously been scored. They compared and discussed the scores for the sample essays to gain an understanding of how the guidelines were to be applied. The faculty consultants were given folders of photocopied essays that were prescored and were asked to use the scoring guidelines to assign scores to the essays. They reviewed their scores and, when there were disagreements, engaged in discussions to compare their rationales for assigning their scores. The process was repeated with additional folders of essays until each faculty consultant demonstrated proficiency in his/her use of the scoring guidelines. (See "Scoring the Free-Response Section" from the *AP Technical Manual* at http://www.collegeboard.com/ap/techman/chap3/scorefc.htm for a more detailed description of the processes of creating scoring guidelines and training faculty consultants to use the scoring guidelines. See also, "Scoring the Exams" from the Released Exam—*1999 AP English Literature and Composition*, The College Board, 1999a, pp. 2–4.)

After the students' essays were scored, their AP grades were determined. For each student, a multiple-choice score and a free-response score were calculated. The two section scores were weighted, and a composite score was calculated. (Appendix, Figure A1, contains a scoring worksheet, showing how the composite scores were derived.) The composite score range was then converted to the AP grade scale, dividing the composite scores into five groups. (For details of how these processes were carried out, see "Calculating the Composite Score" from the *AP Technical Manual* at http://www.collegeboard.com/techman/chap3/composite.htm and "Calculating the AP Grade" at http://www.collegeboard.com/ap/techman/chap3/grade.htm. Also, see "How the AP Grades are Determined" from the *Released Exam—1999 AP English Literature and Composition*, The College Board, 1999a, pp. 71–72.)

## *Procedure*

Data analyses were conducted using the FACETS computer program (Linacre, 1998) and SAS. Severity and other faculty consultant errors were examined using the procedures proposed by Engelhard (1994) and Myford and Wolfe (2001). Before these psychometric indicators of faculty consultant errors were estimated, three approaches for calibrating and linking faculty consultants were explored. Descriptions of those approaches are provided below.

Two views of the assessment design that undergirds the AP ELC assessment system are shown in Table 3. The top panel of Table 3 illustrates one view of the assessment design. From this perspective, the assessment is conceptualized as a two-facet design—students crossed with questions (both multiple-choice and free-response), with faculty consultant effects ignored in the assessment system. As this design shows, each student responded to the 55 multiple-choice questions and wrote three essays, one for each of the three free-response questions. Under this two-facet design, faculty consultants are assumed to be exchangeable—in fact, the goal of the faculty consultant quality control procedures currently in place is to try to achieve exchangeability.

The bottom panel of Table 3 illustrates another possible view of the assessment design. From this perspective, the assessment is conceptualized as a three-facet design—students crossed with questions, and faculty consultants nested within the three free-response questions. Each student responded to the 55 multiple-choice questions, but only three faculty consultants out of 612 scored the student's essays. The bottom panel of Table 3 highlights the incomplete and non-linked nature of the AP ELC assessment system. This non-linked design makes the direct calibration and evaluation of faculty consultant effects impossible without additional assumptions. In the AP ELC assessment system each student writes essays for three separate free-response questions. A single faculty consultant rates each essay, and no faculty consultant rates essays written for more than one free-response question (i.e., in terms of design constraints, faculty consultants are nested within the three free-response questions, and there is no overlap among them). The lack of connectivity between faculty consultants across free-response questions leads to ambiguity in the calibration of faculty consultant severity and free-response question difficulty. Faculty consultant severity, free-response question difficulty, and level of student achievement are confounded. If the average rating for a faculty consultant is lower than the average rating of other consultants (i.e., the faculty consultant appears to be "severe" in comparison to others), it is not clear whether the faculty consultant tended to assign systematically lower ratings than other faculty consultants, or whether the set of student essays the faculty consultant evaluated tended to be lower in quality than other students' essays.

These types of data collection designs have been called nonlinked assessment networks (Engelhard, 1997). This study considered several approaches for calibrating faculty consultants in nonlinked assessment networks (Engelhard, 1997). The specific approaches and assumptions are as follows:

1. The first approach is to anchor student levels of achievement to have a mean of zero within each

TABLE 3

**The AP English Literature and Composition Assessment System Depicted as a Two-Facet Design (Student Crossed with Question, Ignoring Faculty Consultant)**

| | *Multiple-Choice* | | | *Free-Response* | | |
|---|---|---|---|---|---|---|
| Question: | 1 | ■■■ | 55 | 56 | 57 | 58 |
| Student 1 | □ | □ | □ | □ | □ | □ |
| Student 2 | □ | □ | □ | □ | □ | □ |
| Student 3 | □ | □ | □ | □ | □ | □ |
| Student 4 | □ | □ | □ | □ | □ | □ |
| Student 5 | □ | □ | □ | □ | □ | □ |

**The AP English Literature and Composition Assessment System Depicted as a Three-Facet Design (Student Crossed with Question, Faculty Consultant)**

| | *Multiple-Choice* | | | *Free-Response* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Question: | 1 | ■■■ | 55 | 56 | | | 57 | | | 58 | | |
| Faculty Consultant: | ■ | ■ | ■ | 1 | ■■■ | 203 | 204 | ■■■ | 402 | 403 | ■■■ | 612 |
| Student 1 | □ | □ | □ | □ | | | | □ | | | □ | |
| Student 2 | □ | □ | □ | | □ | | □ | | | | □ | |
| Student 3 | □ | □ | □ | | | □ | | □ | | | □ | |
| Student 4 | □ | □ | □ | | □ | | | □ | | □ | | |
| Student 5 | □ | □ | □ | | □ | | | | □ | | | □ |

faculty consultant. This reflects the assumption that students are randomly assigned to faculty consultants, and that these randomly assigned student groups have equivalent average achievement. After the faculty consultants are calibrated, then student ratings on the essays can be estimated with the faculty consultant severities anchored.

2. A second approach is to use multiple-imputation technology to directly estimate the ratings that would be obtained across faculty consultants or pairs of faculty consultants as if they were directly linked (Little and Rubin, 1987; Rubin, 1987). These imputed ratings can then be analyzed as a completely crossed design using the FACETS computer program. This design imputes ratings into the empty cells shown in the bottom panel of Table 3.

3. A third approach is to calibrate the multiple-choice questions (MCQs), free-response questions, and faculty consultants simultaneously. This approach links the faculty consultants through the multiple-choice questions. It assumes that the multiple-choice and free-response questions can be calibrated onto a unidimensional scale. The simultaneous calibration of the MCQs and faculty consultants is congruent with the use of MCQs to monitor faculty consultant behavior through the Reader Management System.

The large number faculty consultants (600+), and the large number of students (170,000 +) make it impractical to develop an operational AP ELC assessment system based on the first two approaches. The anchoring involved in Approach One involves tracking over 170,000 students, and the utility of this approach for calibrating more than 600 faculty consultants simply does not support the potential efforts and costs needed for this approach. Approach Two would involve the imputation of at least five values for each of the 600 + faculty consultants for each of the 170,000 students. As with Approach One, we cannot recommend this approach for the calibration of faculty consultants because of the sheer volume, computational effort, and associated costs. Approach Three was applied and evaluated in this study. As will be seen, the simultaneous calibration of students, questions (combining multiple-choice and free-response), and faculty consultants provides a defensible psychometric framework. The framework appears to offer a highly promising approach for examining and evaluating faculty consultant effects and potential biases within the context of the AP assessment system.[4]

Once the faculty consultants were linked, then the following procedures were used to address each research question. In order to answer Question 1, the parameters for the model in Equation 1 were estimated using the FACETS computer program (Linacre, 1998).

Question 2 was addressed by adding interaction effects to the model in Equation 1. The student characteristics that were examined are gender, race/ethnicity, and best language. In order to estimate these interactions, a facet term and an interaction term for each student characteristic were added to Equation 1. For example, shown below is the many-faceted measurement model we used to investigate the faculty consultant x student gender interaction.

$$ln[P_{nijk} / P_{nijk-1}] = \Theta_n - \xi_i - \alpha_j - \gamma_l - \alpha_j\gamma_l - \tau_k ,$$

**where**

$P_{nijk}$ = probability of student *n* of gender group *l* receiving a rating of *k* on question *i* from faculty consultant *j*,

$P_{nijk-1}$ = probability of student *n* of gender group *l* receiving a rating of *k – 1* on question *i* from faculty consultant *j*,

$\Theta_n$ = English achievement for student *n*,

$\xi_i$ = difficulty of question *i* (including multiple-choice and free-response questions),

$\alpha_j$ = severity of faculty consultant *j*,

$\gamma_l$ = student gender group *l*,

$\alpha_j\gamma_l$ = interaction between severity of faculty consultant *j* and gender group *l*, and

$\tau_k$ = difficulty of receiving a rating of *k* relative to a rating of *k* – 1.

[4] For example, in 1999, the overall correlation between performance on the multiple-choice section of the exam and performance on the free-response section was .521. The correlations of performance on the multiple-choice section with performance on each of the three separate essays were as follows: poetry analysis essay, .365; prose analysis, .444; and analytical-expository essay, .403 (The College Board, 1999). Some critics of our approach to establishing connectivity might argue that linking raters through the multiple-choice questions unfairly disadvantages those students who perform poorly on the multiple-choice portion of the exam but well on the free-response portion. Since the students' responses to the multiple-choice and free-response questions are combined to generate the overall composite score, our approach is consistent with current AP scoring practices. The correlations of performance on the multiple-choice and free-response questions are comparable to what would be expected for correlations between subtest scores on multiple-choice tests. Our results indicate that the overall composite scores are consistent, and that discrepancies between individual student responses on the two exam sections can be detected through the student fit statistics. As will be shown later, the number of misfitting students is quite low.

Three separate FACETS models containing interaction terms were estimated–one for each student characteristic. In essence, Question 2 explores the relationship between faculty consultant severity and the three student background characteristics. These analyses represent differential facet functioning (DFF), and can be viewed as "bias" analyses of faculty consultant performance across various subgroups of students. The approach used is analogous to traditional analyses of differential item functioning (DIF) for multiple-choice questions across subgroups. Question 2 examines whether or not faculty consultant severity is invariant across subgroups: Is the ordering of faculty consultants by severity comparable across subgroups, or are some faculty consultants more severe when they rate students from particular subgroups? We refer to these analyses as differential faculty consultant functioning (DFCF) analyses.

In order to address Question 3, the estimates of student achievement, $\theta$, obtained from Equation 1 were used to generate ratings for the essays that were adjusted for faculty consultant severity. The theta scale was converted back to the reporting scale used for the AP ELC Exams (1 to 9 for the free-response questions). These adjusted ratings were then compared to the observed ratings obtained from the faculty consultants. The mapping of the composite scores (multiple-choice and free-response) onto the final AP grading scale (1 = no recommendation, 2 = possibly qualified,[3] = qualified, 4 = well qualified, 5 = extremely well qualified) was also compared. This methodology for estimating the potential impact of faculty consultant severity was used by Engelhard, Myford, and Cline (2000) to explore assessor severity effects on the performance assessments developed by the National Board for Professional Teaching Standards Examinations

For Question 4, the analyses conducted for Question 3 were summarized for student race/ethnicity subgroups, gender subgroups, and for language subgroups. The two achievement estimates, one adjusted for faculty consultant severity and the other unadjusted, were used to generate expected ratings that were compared within each subgroup. These analyses explored whether or not faculty consultant severity differentially impacted the final AP grades assigned to various student subgroups.

# Results

The results section is divided into several subsections. The first subsection presents selected results from the FACETS analyses of the 1999 AP ELC data. We discuss the various types of practically useful information that a FACETS analysis can provide about individual students, faculty consultants, questions, and rating scales making up an AP assessment. Our intent in this subsection is to show how a many-faceted Rasch measurement (MFRM) approach can be used for quality control monitoring of AP assessments and to pinpoint the potential benefits that could be derived from adopting such an approach. The second subsection looks at the question of whether faculty consultants were differentially severe, exhibiting varying levels of severity depending upon student background characteristics (i.e., gender, race/ethnicity, and best language). The third subsection explores the impact of differences in faculty consultant severity on AP composite scores and grades. We also look at the impact of those severity differences on the AP composite scores and grades for various student subgroups (i.e., is there a systematic relationship between faculty consultant severity and subgroup membership?). The final subsection summarizes the results of our analyses in terms of the specific research questions we posed.

## *FACETS Analyses*

### Variable Map

Figure 2 displays a variable map representing the calibrations of the students, questions, faculty consultants, and the 9-point rating scale for the 1999 AP ELC assessment. Table 4 provides various summary statistics from the FACETS analysis for the three facets (i.e., measures, fit statistics [INFIT and OUTFIT], and separation statistics). The FACETS computer program calibrates the faculty consultants, students, questions, and the rating scale so that all facets are positioned on the same scale, creating a single frame of reference for interpreting the results from the analysis. That scale is in log-odds units, or logits, which, under the model, constitute an equal-interval scale. The first column of Figure 2 shows the logit scale. Having a single frame of reference for all the facets of the assessment process facilitates comparisons within and between the various facets of the analysis. This variable map also indicates the expected ratings that are most likely at various points on the AP ELC calibrated scale. The last column of Figure 2 maps the AP ELC 9-point scale to the equal-interval logit scale that FACETS uses. The horizontal dashed lines in the

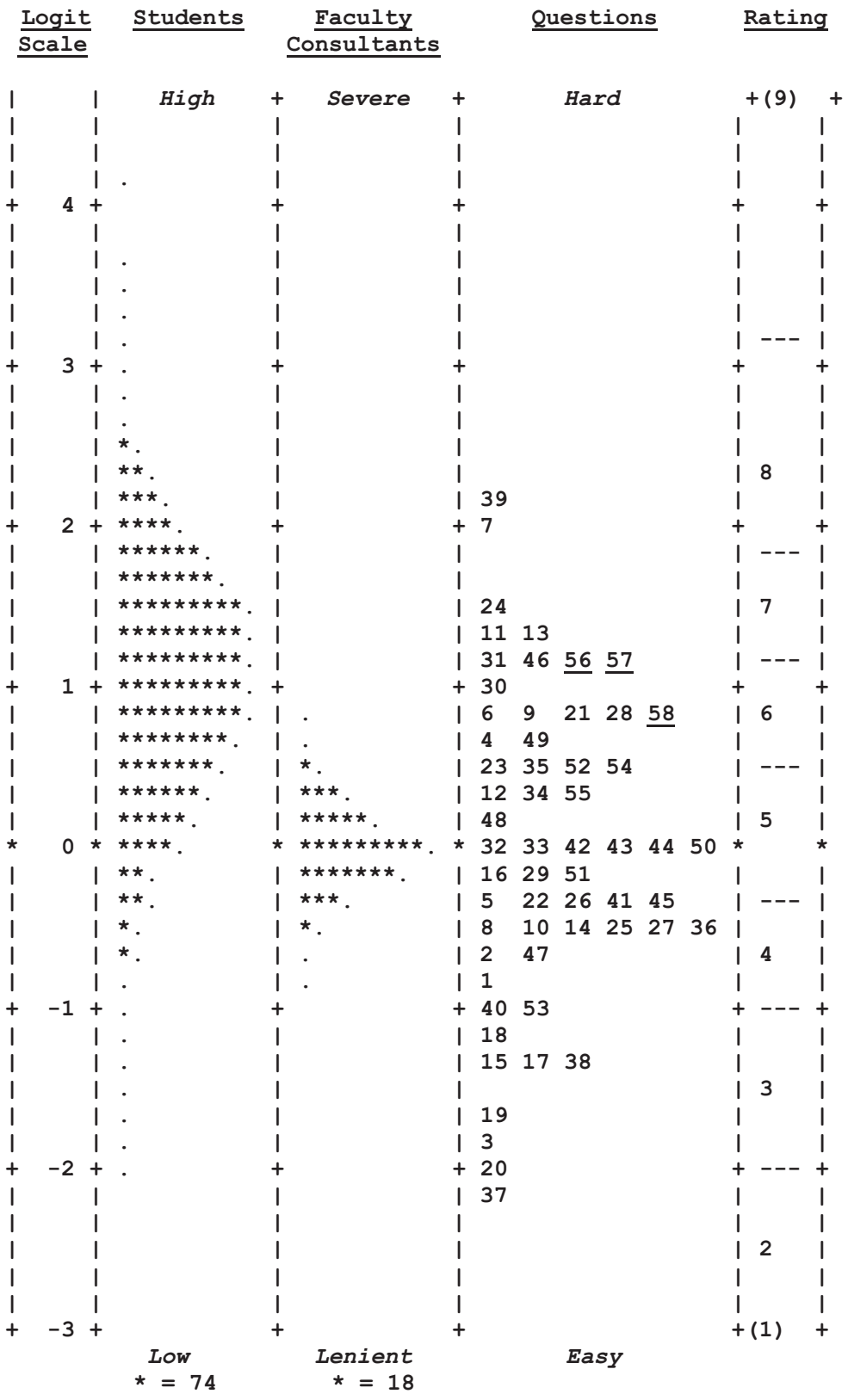


**Figure 2.** Variable map for the 1999 AP English Literature and Composition Assessment.

Note: Free-response questions are underlined.

TABLE 4

**Summary Statistics by Facet (Students, Questions, Faculty Consultants)**

| | *Students (Achievement)* | *Questions (Difficulty)* | *Faculty Consultants (Severity)* | | | |
|---|---|---|---|---|---|---|
| | | | *Total* | *Question 56* | *Question 57* | *Question 58* |
| *Measures* | | | | | | |
| Mean | 1.00 | .00 | .00 | .08 | .06 | -.14 |
| SD | .77 | .97 | .26 | .27 | .24 | .26 |
| N | 8,642 | 58 | 605 | 200 | 195 | 210 |
| *OUTFIT* | | | | | | |
| Mean | 1.0 | 1.0 | 1.2 | 1.1 | 1.1 | 1.2 |
| SD | .3 | .1 | .4 | .3 | .4 | .4 |
| *INFIT* | | | | | | |
| Mean | 1.0 | 1.0 | 1.1 | 1.1 | 1.1 | 1.2 |
| SD | .3 | .1 | .4 | .3 | .4 | .4 |
| *Separation Statistics* | | | | | | |
| Separation Index | 2.50 | 33.96 | 1.57 | 1.67 | 1.46 | 1.56 |
| Reliability of Separation | .86 | .99 | .71 | .71 | .68 | .71 |
| Fixed Chi-Square Statistic | 58580.9* | 72684.9* | 2203.7* | 795.4* | 606.9* | 799.2* |
| df | 8641 | 57 | 604 | 199 | 194 | 209 |

* $p < .01$

last column are positioned at the point at which the expected, i.e., average over the long run, rating is "lower rating + 0.5." The mapping makes evident the approximate equal intervals in the AP ELC scale. (Scale points 1 and 9, which would indicate perfect extreme ratings, are off the scale of this figure and thus are not shown.)

The second column of the variable map displays the student measures of English achievement. Higher-scoring students appear at the top of the column, while lower-scoring students appear at the bottom. Each star represents 74 students, and a dot represents 1–73 students. The student achievement measures range from –1.96 logits to 4.10 logits, with 50 percent of the student measures between 0.50 and 1.53 logits ($M$ = 1.00, $SD$ = 0.77, $N$ = 8,642).

The third column compares the faculty consultants in terms of the level of severity or leniency each employed when rating students' essays. In this column, each star represents 18 faculty consultants, and a dot represents 1–17 faculty consultants. More severe faculty consultants appear at the top of the column, while more lenient faculty consultants appear lower in the column. The harshest faculty consultant had a severity measure of 0.87 logits, while the most lenient faculty consultant had a severity measure of -0.76 logits ($M$ = 0.00, $SD$ = 0.26, $N$ = 605). Fifty percent of the faculty consultant severity measures were between -0.17 and 0.16 logits. The variable map shows that the students differ more in their levels of English achievement than faculty consultants differ in their levels of severity. The distribution of faculty consultant severity measures is much narrower than the distribution of student achievement measures. The faculty consultant severity measures show a 1.63-logit spread, while the student achievement measures show a 6.06-logit spread.

The questions are shown in the fourth column of the variable map. Questions appearing higher in the column were more difficult for students than questions appearing lower in the column. Questions 1 to 55 were multiple-choice questions, while Questions 56, 57, and 58 were the three free-response questions. (The free-response questions are underlined on the variable map.) The variable map shows that the average level of achievement for the students ($M$ = 1.00) is higher than the average question difficulty ($M$ = 0.00). It should also be noted that the three free-response questions are more difficult ($M$ = 1.07) than the multiple-choice questions ($M$ = -0.06). The question difficulty measures show a 4.22-logit spread, while the student achievement measures show a 6.06-logit spread. The question difficulties range from –2.14 to 2.08 logits ($M$ = 0.00, $SD$ = 0.97, $N$ = 58).

## Rating Scale

Table 5 presents information regarding the functioning of the 9-category rating scale that faculty consultants employed to evaluate students' essays. The faculty consultants used all the categories, with approximately 90 percent of the ratings assigned in categories 3 through 7. Table 5 reports an "average measure" for each rating category. If the rating scale is functioning as intended, then the average measures will increase in magnitude as the rating scale categories increase. When this pattern is borne out in the data, the results suggest that students with higher ratings are indeed exhibiting more of the

TABLE 5

**Scale Category Statistics**

| | Count | | Category Calibrations | | Fit | |
|---|---|---|---|---|---|---|
| *Category* | *Freq.* | *%* | *Threshold* | *SE* | *Average Measure* | *OUTFIT MNSQ* |
| 1 | 348 | 1 | | | -1.40 | 1.2 |
| 2 | 1,349 | 5 | -2.69 | .06 | -1.01 | 1.2 |
| 3 | 3,444 | 13 | -1.86 | .03 | -.67 | 1.1 |
| 4 | 6,180 | 24 | -1.11 | .02 | -.31 | 1.1 |
| 5 | 5,904 | 23 | -.11 | .02 | .01 | 1.1 |
| 6 | 4,915 | 19 | .37 | .02 | .31 | 1.2 |
| 7 | 2,523 | 10 | 1.18 | .02 | .59 | 1.2 |
| 8 | 990 | 4 | 1.77 | .03 | .95 | 1.1 |
| 9 | 273 | 1 | 2.45 | .06 | 1.24 | 1.1 |

construct being measured (i.e., English achievement) than students with lower ratings. Therefore, the intentions of those who designed the rating scale are being fulfilled, according to Linacre (1999). As Table 5 shows, the average measures of the students increased as the rating categories increased. Indeed, the increases in the average measures are nearly uniform (i.e., 0.33, + or – 0.05).

The category thresholds are another useful indicator of whether a rating scale is working as intended. A threshold denotes the point at which the probability curves of two adjacent rating scale categories cross (Linacre, 1999). Thus, the rating scale category threshold represents the point at which the probability is 50 percent of a student being rated in one or the other of these two categories, given that the student is in one of them (Andrich, 1998). An important requirement for maintaining score interpretability is that the scale category thresholds advance monotonically (Andrich, 1998). If the scale category thresholds do not increase in value, then these disordered thresholds can muddle the interpretations of the categories, since one or more of the categories are never most probable to be assigned. Table 5 shows a clear progression of the scale category thresholds from–2.69 logits (i.e., the threshold between categories 1 to 2) to 2.45 logits (i.e., the threshold between categories 8 to 9).

Table 5 presents a third useful indicator of rating scale functionality—an OUTFIT mean-square statistic for each rating category. For each rating scale category, FACETS computes the average student achievement measure (i.e., the "observed" measure) and an "expected" student achievement measure (i.e., the student measure the model would predict for that rating category if the data were to fit the model). When the observed and expected student achievement measures are close, then the OUTFIT mean-square statistic for the rating category will be near the expected value of 1.0. The greater the discrepancy between the observed and expected measures, the larger the OUTFIT mean-square statistic will be. For a given rating category, an OUTFIT mean-square statistic of 1.0, the expected value according to the model, implies that the ratings, on average, are each contributing one unit of statistical information to the measurement process, as they are intended to do. Values less than 0.5 suggest that the ratings are overly predictable from each other. This does not degrade the student achievement measures themselves, but may bias upward computations of reliability and separation. (This is known as the "attenuation paradox" in classical psychometrics.) Values greater than 1.5 suggest that there is a noticeable unexpected component in the ratings, while values greater than 2.0 indicate that there is more unexplained variability (i.e., noise) in the ratings than statistical information. This may or may not bias the student achievement measures depending on whether the noise pattern is symmetrical or asymmetrical. In any case, it lessens the precision of the achievement measures (J.M. Linacre, personal communication, July 5, 2000).

The OUTFIT mean-square statistics shown in Table 5 for the various rating categories are all near the expected value of 1.0. It is important to note that our presentation of results regarding the functioning of the AP ELC rating scale does not examine how the scale functions for each individual free-response question. Rather, we looked at how the scale performed across the three free-response questions (that is, as a "common" scale). Based on the results of our analyses, there is reason to believe that the faculty consultants used the 9-point rating scale in a similar fashion across the three free-response questions.

## Students

As shown in Table 4, the 8,642 students in the 5 percent sample can be separated into about two and a half statistically distinct levels of English achievement (i.e., the student separation index is 2.50). This variability is statistically significant, $\chi^2$ (8641, $N$ = 8642) = 58,580.9 $p < .01$, and the reliability of separation for the students is quite high ($R$ = .86). The reliability of separation for students is comparable to the reliability estimate

($R$ = .850) in the 1998 AP ELC reader reliability study (Maneckshana, Morgan, and Batleman, 1999). The standard error of measurement is 0.28 logits on average for the students based on the "model" RMSE, based on the assumption that all randomness in the data accords with that predicted by the model.

Both the student INFIT ($M$ = 1.0, $SD$ = 0.3) and OUTFIT mean-square summary statistics ($M$ = 1.0, $SD$ = 0.3) indicate overall good fit of data to the model. Out of the 8,642 students, approximately 90 percent of the students had OUTFIT mean-square statistics between 0.6 and 1.5. There were only 352 students (4.1 percent) who had OUTFIT mean-square statistics greater than 1.5.[5] It should be noted that there were 514 students (5.9 percent) who had OUTFIT mean-square statistics less than 0.6; within the context of rater-mediated assessments, it is not recommended that these low values for students be investigated as misfitting cases (Engelhard, 1994).[6] Given the large number of students, this is not an unusual number of unlikely rating patterns. This finding provides support for the inference that invariant measurement has been accomplished within the framework of the many-faceted Rasch measurement model.

Quality control tables and charts can be constructed for those students exhibiting unusual score profiles to identify the specific multiple-choice questions that a student unexpectedly answered incorrectly (or correctly). These tables and charts also isolate the specific ratings the student received that were highly unexpected or surprising, given the faculty consultant's level of severity, the student's performance on the multiple-choice questions, and the other ratings that the student received on the free-response questions. Tables 6, 7, and 8 present illustrative quality control tables for Students 2508, 6019, and 1851 respectively. For this study, we adopted an upper-control limit for the student fit mean-square statistics of 1.5 and a lower-control limit of 0.6.[7] The fit mean-square statistics for Student 2508 are above the upper-control limit, within the limits for Student 6019, and at the lower-control limit for Student 1851. The profile of ratings for Student 2508 shows more variability than the model would expect, while the rating profile for Student 1851 shows less variability than expected. As a point of contrast, the rating profile for Student 6019 shows an expected amount of variability.

The diagnostic data for these three students are presented graphically in Figures 3, 4, and 5. For questions 1–55, a value of 1 in the "Observed" column indicates that the student answered the multiple-choice question correctly, while a value of 0 indicates that the student answered the question incorrectly. For questions 56–58, the value appearing in the "Observed" column is the rating the student received on the essay composed for that free-response question. $Z$-scores greater than zero indicate observed responses and ratings higher than expected based on the MFRM model, while $z$-scores less than zero reflect observed values that are lower than expected. As a general guideline, values above + 2.0 and below –2.0 can be viewed as reflecting statistically significant differences between observed and expected values. An independent review to confirm the AP grade might be warranted for a student having a highly unusual score profile (i.e., one or more unexpected ratings on the free-response questions and/or a number of unexpectedly incorrect [or correct] responses on the multiple-choice questions).

For example, as shown in Table 6 and Figure 2, Student 2508 (INFIT MNSQ = 2.0, OUTFIT MNSQ = 2.3) has a score profile that shows a great deal of variability, more variability in performance across the questions than expected. Student 2508 unexpectedly answered nine multiple-choice questions incorrectly (1, 2, 5, 15, 16, 23, 34, 35, and 37). Given the student's overall level of achievement and the difficulties of these particular questions, the measurement model would have expected the student to answer these questions correctly. This student also received an unexpectedly high rating from Faculty Consultant 666 on the third free-response question (Question 58, $z$-score = 2.34). The faculty consultant gave the student's essay an 8, but the model predicted that the consultant should give the essay a 5, given the faculty consultant's level of severity and the difficulty of the question. Faculty Consultant 666 was one of the consultants with average severity (severity measure = -0.07 logits), so it is somewhat surprising that the

[5] The expectation for this statistic is 1; the range is 0 to infinity. The higher the OUTFIT mean-square statistic, the greater the variability in the student's score profile, even when faculty consultant severity is taken into account. An OUTFIT mean-square statistic greater than 1 indicates more than typical variation in the ratings on the free-response questions and/or scores on the multiple-choice questions (that is, a set of scores in which one or more are highly unexpected and thus don't seem to "fit" with the others).

[6] An OUTFIT mean-square statistic less than 1 indicates little variation in the student's score profile (i.e., overly predictable performance across the set of 58 questions, too little variation in the scores).

[7] No hard-and-fast rules exist for establishing upper- and lower-control limits for student fit mean-square statistics. Some testing programs use an upper-control limit of 2 or 3 and a lower-control limit of 0.5, but more stringent limits might be set if the program desired to reduce significantly variability within the assessment system. The more extreme the fit mean-square statistic, the greater potential gains for improving the assessment system.

TABLE 6

**Quality Control Table for Student 2508 (INFIT MNSQ = 2.0, OUTFIT MNSQ = 2.3)**

| *Index* | *Student ID* | *Student Gender* | *Student Race/Ethnicity* | *Student Best Language* | *Question* | *Faculty Consultant* | *Observed Rating* | *Expected Rating* | *Residual* | *Z-score* |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2508 | 1 | 7 | 3 | 1 | . | 0 | 0.87 | -0.87 | -2.56 |
| 2 | 2508 | 1 | 7 | 3 | 2 | . | 0 | 0.86 | -0.86 | -2.45 |
| 3 | 2508 | 1 | 7 | 3 | 3 | . | 1 | 0.95 | 0.05 | 0.23 |
| 4 | 2508 | 1 | 7 | 3 | 4 |  | 0 | 0.60 | -0.60 | -1.22 |
| 5 | 2508 | 1 | 7 | 3 | 5 | . | 0 | 0.80 | -0.80 | -2.03 |
| 6 | 2508 | 1 | 7 | 3 | 6 | . | 1 | 0.58 | 0.42 | 0.86 |
| 7 | 2508 | 1 | 7 | 3 | 7 | . | 0 | 0.30 | -0.30 | -0.65 |
| 8 | 2508 | 1 | 7 | 3 | 8 | . | 1 | 0.83 | 0.17 | 0.46 |
| 9 | 2508 | 1 | 7 | 3 | 9 | . | 1 | 0.57 | 0.43 | 0.86 |
| 10 | 2508 | 1 | 7 | 3 | 12 | . | 0 | 0.69 | -0.69 | -1.50 |
| 11 | 2508 | 1 | 7 | 3 | 13 | . | 1 | 0.42 | 0.58 | 1.18 |
| 12 | 2508 | 1 | 7 | 3 | 14 | . | 1 | 0.83 | 0.17 | 0.45 |
| 13 | 2508 | 1 | 7 | 3 | 15 | . | 1 | 0.92 | 0.08 | 0.30 |
| 14 | 2508 | 1 | 7 | 3 | 16 | . | 0 | 0.79 | -0.79 | -1.93 |
| 15 | 2508 | 1 | 7 | 3 | 17 | . | 0 | 0.92 | -0.92 | -3.42 |
| 16 | 2508 | 1 | 7 | 3 | 18 | . | 0 | 0.90 | -0.90 | -3.04 |
| 17 | 2508 | 1 | 7 | 3 | 19 | . | 1 | 0.94 | 0.06 | 0.26 |
| 18 | 2508 | 1 | 7 | 3 | 20 | . | 1 | 0.96 | 0.04 | 0.21 |
| 19 | 2508 | 1 | 7 | 3 | 21 | . | 0 | 0.54 | -0.54 | -1.08 |
| 20 | 2508 | 1 | 7 | 3 | 22 | . | 0 | 0.79 | -0.79 | -1.93 |
| 21 | 2508 | 1 | 7 | 3 | 23 | . | 1 | 0.62 | 0.38 | 0.78 |
| 22 | 2508 | 1 | 7 | 3 | 24 | . | 0 | 0.41 | -0.41 | -0.84 |
| 23 | 2508 | 1 | 7 | 3 | 25 | . | 0 | 0.83 | -0.83 | -2.23 |
| 24 | 2508 | 1 | 7 | 3 | 26 | . | 1 | 0.80 | 0.20 | 0.51 |
| 25 | 2508 | 1 | 7 | 3 | 27 | . | 1 | 0.83 | 0.17 | 0.45 |
| 26 | 2508 | 1 | 7 | 3 | 28 | . | 1 | 0.58 | 0.42 | 0.85 |
| 27 | 2508 | 1 | 7 | 3 | 29 | . | 1 | 0.76 | 0.24 | 0.56 |
| 28 | 2508 | 1 | 7 | 3 | 30 | . | 1 | 0.50 | 0.50 | 1.01 |
| 29 | 2508 | 1 | 7 | 3 | 31 | . | 0 | 0.48 | -0.48 | -0.95 |
| 30 | 2508 | 1 | 7 | 3 | 32 | . | 1 | 0.73 | 0.27 | 0.60 |
| 31 | 2508 | 1 | 7 | 3 | 33 | . | 1 | 0.74 | 0.26 | 0.59 |
| 32 | 2508 | 1 | 7 | 3 | 35 | . | 1 | 0.65 | 0.35 | 0.73 |
| 33 | 2508 | 1 | 7 | 3 | 36 | . | 1 | 0.82 | 0.18 | 0.47 |
| 34 | 2508 | 1 | 7 | 3 | 37 | . | 0 | 0.96 | -0.96 | -4.97 |
| 35 | 2508 | 1 | 7 | 3 | 38 | . | 0 | 0.92 | -0.92 | -3.36 |
| 36 | 2508 | 1 | 7 | 3 | 39 | . | 0 | 0.27 | -0.27 | -0.60 |
| 37 | 2508 | 1 | 7 | 3 | 41 | . | 0 | 0.80 | -0.80 | -2.01 |
| 38 | 2508 | 1 | 7 | 3 | 42 | . | 1 | 0.75 | 0.25 | 0.58 |
| 39 | 2508 | 1 | 7 | 3 | 43 | . | 1 | 0.74 | 0.26 | 0.60 |
| 40 | 2508 | 1 | 7 | 3 | 44 | . | 1 | 0.73 | 0.27 | 0.60 |
| 41 | 2508 | 1 | 7 | 3 | 45 | . | 1 | 0.81 | 0.19 | 0.49 |
| 42 | 2508 | 1 | 7 | 3 | 46 | . | 0 | 0.49 | -0.49 | -0.97 |
| 43 | 2508 | 1 | 7 | 3 | 47 | . | 1 | 0.86 | 0.14 | 0.41 |
| 44 | 2508 | 1 | 7 | 3 | 48 | . | 1 | 0.72 | 0.28 | 0.62 |
| 45 | 2508 | 1 | 7 | 3 | 49 | . | 1 | 0.58 | 0.42 | 0.84 |
| 46 | 2508 | 1 | 7 | 3 | 50 | . | 1 | 0.75 | 0.25 | 0.58 |
| 47 | 2508 | 1 | 7 | 3 | 51 | . | 1 | 0.78 | 0.22 | 0.53 |
| 48 | 2508 | 1 | 7 | 3 | 52 | . | 1 | 0.65 | 0.35 | 0.74 |
| 49 | 2508 | 1 | 7 | 3 | 53 | . | 1 | 0.89 | 0.11 | 0.34 |
| 50 | 2508 | 1 | 7 | 3 | 54 | . | 1 | 0.65 | 0.35 | 0.73 |
| 51 | 2508 | 1 | 7 | 3 | 55 | . | 0 | 0.66 | -0.66 | -1.41 |
| 52 | 2508 | 1 | 7 | 3 | 56 | 388 | 7 | 4.81 | 2.19 | 1.87 |
| 53 | 2508 | 1 | 7 | 3 | 57 | 205 | 6 | 5.17 | 0.83 | 0.70 |
| 54 | 2508 | 1 | 7 | 3 | 58 | 666 | 8 | 5.22 | 2.78 | 2.34 |

Note. This student did not answer Questions 10, 11, 34, and 40.

TABLE 7

**Quality Control Table for Student 6019 (INFIT MNSQ = 0.9, OUTFIT MNSQ = 1.0)**

| *Index* | *Student ID* | *Student Gender* | *Student Race/Ethnicity* | *Student Best Language* | *Question* | *Faculty Consultant* | *Observed Rating* | *Expected Rating* | *Residual* | *Z-score* |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 6019 | 1 | 7 | 1 | 1 | . | 1 | 0.86 | 0.14 | 0.40 |
| 2 | 6019 | 1 | 7 | 1 | 2 | . | 1 | 0.85 | 0.15 | 0.41 |
| 3 | 6019 | 1 | 7 | 1 | 3 | . | 1 | 0.95 | 0.05 | 0.23 |
| 4 | 6019 | 1 | 7 | 1 | 4 | . | 0 | 0.59 | -0.59 | -1.20 |
| 5 | 6019 | 1 | 7 | 1 | 5 | . | 1 | 0.80 | 0.20 | 0.50 |
| 6 | 6019 | 1 | 7 | 1 | 6 | . | 1 | 0.57 | 0.43 | 0.87 |
| 7 | 6019 | 1 | 7 | 1 | 7 | . | 0 | 0.29 | -0.29 | -0.64 |
| 8 | 6019 | 1 | 7 | 1 | 8 | . | 1 | 0.82 | 0.18 | 0.47 |
| 9 | 6019 | 1 | 7 | 1 | 9 | . | 1 | 0.57 | 0.43 | 0.87 |
| 10 | 6019 | 1 | 7 | 1 | 10 | . | 1 | 0.83 | 0.17 | 0.45 |
| 11 | 6019 | 1 | 7 | 1 | 11 | . | 0 | 0.41 | -0.41 | -0.84 |
| 12 | 6019 | 1 | 7 | 1 | 12 | . | 1 | 0.69 | 0.31 | 0.67 |
| 13 | 6019 | 1 | 7 | 1 | 13 | . | 1 | 0.41 | 0.59 | 1.19 |
| 14 | 6019 | 1 | 7 | 1 | 14 | . | 1 | 0.83 | 0.17 | 0.45 |
| 15 | 6019 | 1 | 7 | 1 | 15 | . | 1 | 0.92 | 0.08 | 0.30 |
| 16 | 6019 | 1 | 7 | 1 | 16 | . | 1 | 0.78 | 0.22 | 0.53 |
| 17 | 6019 | 1 | 7 | 1 | 17 | . | 1 | 0.92 | 0.08 | 0.30 |
| 18 | 6019 | 1 | 7 | 1 | 18 | . | 1 | 0.90 | 0.10 | 0.33 |
| 19 | 6019 | 1 | 7 | 1 | 19 | . | 1 | 0.94 | 0.06 | 0.26 |
| 20 | 6019 | 1 | 7 | 1 | 20 | . | 1 | 0.96 | 0.04 | 0.21 |
| 21 | 6019 | 1 | 7 | 1 | 21 | . | 1 | 0.53 | 0.47 | 0.93 |
| 22 | 6019 | 1 | 7 | 1 | 22 | . | 1 | 0.79 | 0.21 | 0.52 |
| 23 | 6019 | 1 | 7 | 1 | 23 | . | 0 | 0.62 | -0.62 | -1.27 |
| 24 | 6019 | 1 | 7 | 1 | 24 | . | 1 | 0.41 | 0.59 | 1.21 |
| 25 | 6019 | 1 | 7 | 1 | 25 | . | 1 | 0.83 | 0.17 | 0.45 |
| 26 | 6019 | 1 | 7 | 1 | 26 | . | 0 | 0.79 | -0.79 | -1.95 |
| 27 | 6019 | 1 | 7 | 1 | 27 | . | 0 | 0.83 | -0.83 | -2.19 |
| 28 | 6019 | 1 | 7 | 1 | 28 | . | 0 | 0.57 | -0.57 | -1.16 |
| 29 | 6019 | 1 | 7 | 1 | 29 | . | 0 | 0.76 | -0.76 | -1.77 |
| 30 | 6019 | 1 | 7 | 1 | 30 | . | 1 | 0.49 | 0.51 | 1.02 |
| 31 | 6019 | 1 | 7 | 1 | 31 | . | 1 | 0.47 | 0.53 | 1.06 |
| 32 | 6019 | 1 | 7 | 1 | 32 | . | 0 | 0.73 | -0.73 | -1.64 |
| 33 | 6019 | 1 | 7 | 1 | 33 | . | 1 | 0.74 | 0.26 | 0.60 |
| 34 | 6019 | 1 | 7 | 1 | 34 | . | 1 | 0.66 | 0.34 | 0.71 |
| 35 | 6019 | 1 | 7 | 1 | 35 | . | 1 | 0.64 | 0.36 | 0.74 |
| 36 | 6019 | 1 | 7 | 1 | 36 | . | 0 | 0.81 | -0.81 | -2.08 |
| 37 | 6019 | 1 | 7 | 1 | 37 | . | 1 | 0.96 | 0.04 | 0.20 |
| 38 | 6019 | 1 | 7 | 1 | 38 | . | 1 | 0.92 | 0.08 | 0.30 |
| 39 | 6019 | 1 | 7 | 1 | 39 | . | 0 | 0.26 | -0.26 | -0.59 |
| 40 | 6019 | 1 | 7 | 1 | 40 | . | 0 | 0.88 | -0.88 | -2.72 |
| 41 | 6019 | 1 | 7 | 1 | 41 | . | 1 | 0.80 | 0.20 | 0.50 |
| 42 | 6019 | 1 | 7 | 1 | 42 | . | 1 | 0.74 | 0.26 | 0.59 |
| 43 | 6019 | 1 | 7 | 1 | 43 | . | 0 | 0.73 | -0.73 | -1.65 |
| 44 | 6019 | 1 | 7 | 1 | 44 | . | 0 | 0.73 | -0.73 | -1.64 |
| 45 | 6019 | 1 | 7 | 1 | 45 | . | 1 | 0.80 | 0.20 | 0.49 |
| 46 | 6019 | 1 | 7 | 1 | 46 | . | 0 | 0.48 | -0.48 | -0.96 |
| 47 | 6019 | 1 | 7 | 1 | 47 | . | 1 | 0.85 | 0.15 | 0.42 |
| 48 | 6019 | 1 | 7 | 1 | 48 | . | 1 | 0.72 | 0.28 | 0.63 |
| 49 | 6019 | 1 | 7 | 1 | 49 | . | 1 | 0.58 | 0.42 | 0.85 |
| 50 | 6019 | 1 | 7 | 1 | 50 | . | 1 | 0.75 | 0.25 | 0.58 |
| 51 | 6019 | 1 | 7 | 1 | 51 | . | 1 | 0.78 | 0.22 | 0.53 |
| 52 | 6019 | 1 | 7 | 1 | 52 | . | 0 | 0.64 | -0.64 | -1.34 |
| 53 | 6019 | 1 | 7 | 1 | 53 | . | 1 | 0.89 | 0.11 | 0.35 |
| 54 | 6019 | 1 | 7 | 1 | 54 | . | 0 | 0.65 | -0.65 | -1.36 |
| 55 | 6019 | 1 | 7 | 1 | 55 | . | 1 | 0.66 | 0.34 | 0.72 |
| 56 | 6019 | 1 | 7 | 1 | 56 | 650 | 5 | 4.56 | 0.44 | 0.38 |
| 57 | 6019 | 1 | 7 | 1 | 57 | 298 | 4 | 3.97 | 0.03 | 0.02 |
| 58 | 6019 | 1 | 7 | 1 | 58 | 352 | 6 | 5.07 | 0.93 | 0.78 |

TABLE 8

**Quality Control Table for Student 1851 (INFIT MNSQ = 0.7, OUTFIT MNSQ = 0.6)**

| *Index* | *Student ID* | *Student Gender* | *Student Race/Ethnicity* | *Student Best Language* | *Question* | *Faculty Consultant* | *Observed Rating* | *Expected Rating* | *Residual* | *Z-score* |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1851 | 2 | 7 | 1 | 1 | . | 1 | 0.87 | 0.13 | 0.39 |
| 2 | 1851 | 2 | 7 | 1 | 2 | . | 1 | 0.85 | 0.15 | 0.41 |
| 3 | 1851 | 2 | 7 | 1 | 3 | . | 1 | 0.95 | 0.05 | 0.23 |
| 4 | 1851 | 2 | 7 | 1 | 4 | . | 1 | 0.59 | 0.41 | 0.83 |
| 5 | 1851 | 2 | 7 | 1 | 5 | . | 1 | 0.80 | 0.20 | 0.50 |
| 6 | 1851 | 2 | 7 | 1 | 6 | . | 1 | 0.57 | 0.43 | 0.87 |
| 7 | 1851 | 2 | 7 | 1 | 7 | . | 0 | 0.29 | -0.29 | -0.64 |
| 8 | 1851 | 2 | 7 | 1 | 8 | . | 1 | 0.82 | 0.18 | 0.46 |
| 9 | 1851 | 2 | 7 | 1 | 9 | . | 1 | 0.57 | 0.43 | 0.87 |
| 10 | 1851 | 2 | 7 | 1 | 10 | . | 1 | 0.83 | 0.17 | 0.45 |
| 11 | 1851 | 2 | 7 | 1 | 11 | . | 0 | 0.42 | -0.42 | -0.84 |
| 12 | 1851 | 2 | 7 | 1 | 12 | . | 1 | 0.69 | 0.31 | 0.67 |
| 13 | 1851 | 2 | 7 | 1 | 13 | . | 1 | 0.41 | 0.59 | 1.19 |
| 14 | 1851 | 2 | 7 | 1 | 14 | . | 1 | 0.83 | 0.17 | 0.45 |
| 15 | 1851 | 2 | 7 | 1 | 15 | . | 1 | 0.92 | 0.08 | 0.30 |
| 16 | 1851 | 2 | 7 | 1 | 16 | . | 1 | 0.78 | 0.22 | 0.52 |
| 17 | 1851 | 2 | 7 | 1 | 17 | . | 1 | 0.92 | 0.08 | 0.30 |
| 18 | 1851 | 2 | 7 | 1 | 18 | . | 1 | 0.90 | 0.10 | 0.33 |
| 19 | 1851 | 2 | 7 | 1 | 19 | . | 1 | 0.94 | 0.06 | 0.26 |
| 20 | 1851 | 2 | 7 | 1 | 20 | . | 1 | 0.96 | 0.04 | 0.21 |
| 21 | 1851 | 2 | 7 | 1 | 21 | . | 0 | 0.53 | -0.53 | -1.07 |
| 22 | 1851 | 2 | 7 | 1 | 22 | . | 1 | 0.79 | 0.21 | 0.52 |
| 23 | 1851 | 2 | 7 | 1 | 23 | . | 1 | 0.62 | 0.38 | 0.79 |
| 24 | 1851 | 2 | 7 | 1 | 24 | . | 0 | 0.41 | -0.41 | -0.83 |
| 25 | 1851 | 2 | 7 | 1 | 25 | . | 1 | 0.83 | 0.17 | 0.45 |
| 26 | 1851 | 2 | 7 | 1 | 26 | . | 1 | 0.79 | 0.21 | 0.51 |
| 27 | 1851 | 2 | 7 | 1 | 27 | . | 1 | 0.83 | 0.17 | 0.46 |
| 28 | 1851 | 2 | 7 | 1 | 28 | . | 0 | 0.57 | -0.57 | -1.16 |
| 29 | 1851 | 2 | 7 | 1 | 29 | . | 1 | 0.76 | 0.24 | 0.56 |
| 30 | 1851 | 2 | 7 | 1 | 30 | . | 0 | 0.49 | -0.49 | -0.98 |
| 31 | 1851 | 2 | 7 | 1 | 31 | . | 0 | 0.47 | -0.47 | -0.94 |
| 32 | 1851 | 2 | 7 | 1 | 32 | . | 1 | 0.73 | 0.27 | 0.61 |
| 33 | 1851 | 2 | 7 | 1 | 33 | . | 1 | 0.74 | 0.26 | 0.60 |
| 34 | 1851 | 2 | 7 | 1 | 34 | . | 0 | 0.66 | -0.66 | -1.41 |
| 35 | 1851 | 2 | 7 | 1 | 35 | . | 1 | 0.65 | 0.35 | 0.74 |
| 36 | 1851 | 2 | 7 | 1 | 36 | . | 1 | 0.81 | 0.19 | 0.48 |
| 37 | 1851 | 2 | 7 | 1 | 37 | . | 1 | 0.96 | 0.04 | 0.20 |
| 38 | 1851 | 2 | 7 | 1 | 38 | . | 1 | 0.92 | 0.08 | 0.30 |
| 39 | 1851 | 2 | 7 | 1 | 39 | . | 0 | 0.26 | -0.26 | -0.59 |
| 40 | 1851 | 2 | 7 | 1 | 40 | . | 1 | 0.88 | 0.12 | 0.37 |
| 41 | 1851 | 2 | 7 | 1 | 41 | . | 0 | 0.80 | -0.80 | -1.99 |
| 42 | 1851 | 2 | 7 | 1 | 42 | . | 1 | 0.74 | 0.26 | 0.59 |
| 43 | 1851 | 2 | 7 | 1 | 43 | . | 1 | 0.73 | 0.27 | 0.60 |
| 44 | 1851 | 2 | 7 | 1 | 44 | . | 1 | 0.73 | 0.27 | 0.61 |
| 45 | 1851 | 2 | 7 | 1 | 45 | . | 1 | 0.80 | 0.20 | 0.49 |
| 46 | 1851 | 2 | 7 | 1 | 46 | . | 0 | 0.48 | -0.48 | -0.96 |
| 47 | 1851 | 2 | 7 | 1 | 47 | . | 1 | 0.85 | 0.15 | 0.41 |
| 48 | 1851 | 2 | 7 | 1 | 48 | . | 1 | 0.72 | 0.28 | 0.63 |
| 49 | 1851 | 2 | 7 | 1 | 49 | . | 1 | 0.58 | 0.42 | 0.85 |
| 50 | 1851 | 2 | 7 | 1 | 50 | . | 0 | 0.75 | -0.75 | -1.71 |
| 51 | 1851 | 2 | 7 | 1 | 51 | . | 1 | 0.78 | 0.22 | 0.53 |
| 52 | 1851 | 2 | 7 | 1 | 52 | . | 0 | 0.64 | -0.64 | -1.34 |
| 53 | 1851 | 2 | 7 | 1 | 53 | . | 1 | 0.89 | 0.11 | 0.35 |
| 54 | 1851 | 2 | 7 | 1 | 54 | . | 1 | 0.65 | 0.35 | 0.74 |
| 55 | 1851 | 2 | 7 | 1 | 55 | . | 1 | 0.66 | 0.34 | 0.72 |
| 56 | 1851 | 2 | 7 | 1 | 56 | 542 | 4 | 4.69 | -0.69 | -0.59 |
| 57 | 1851 | 2 | 7 | 1 | 57 | 556 | 4 | 4.63 | -0.63 | -0.54 |
| 58 | 1851 | 2 | 7 | 1 | 58 | 654 | 4 | 5.27 | -1.27 | -1.06 |

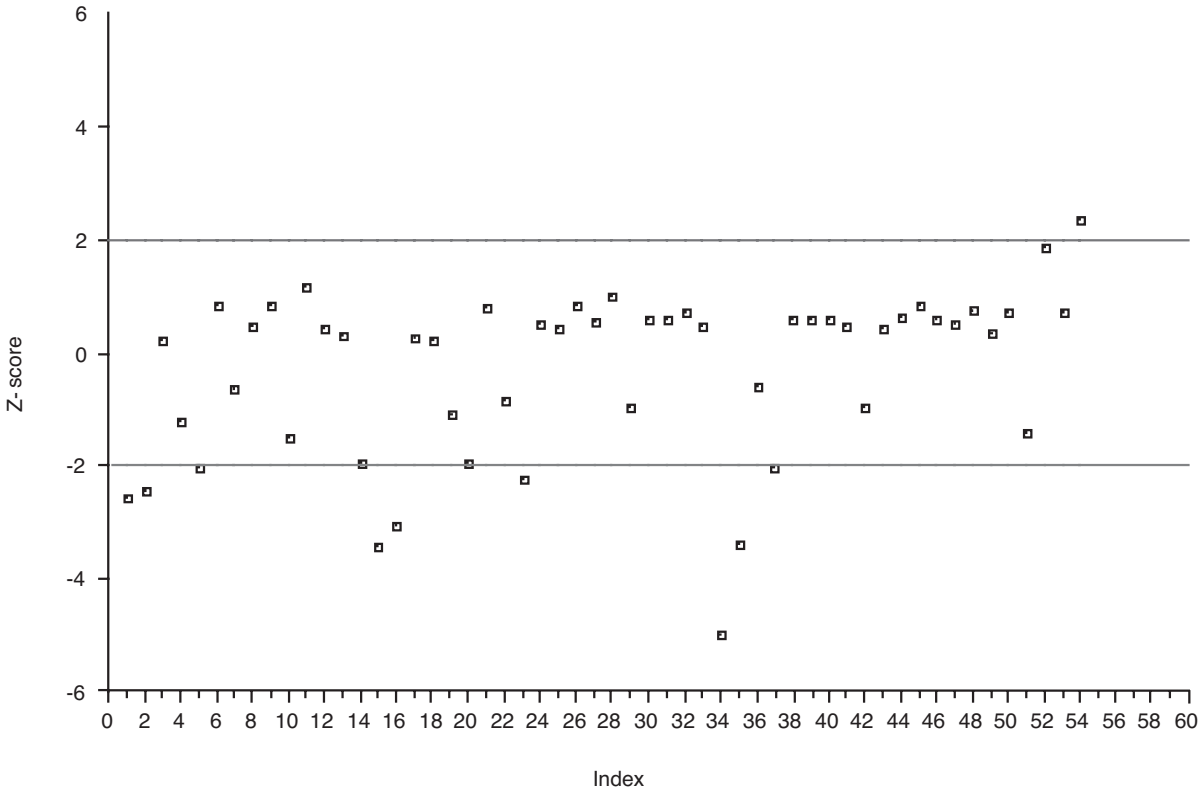


**Figure 3.** Quality control chart for student 2508 (INFIT MNSQ = 2.0, OUTFIT MNSQ = 2.3).

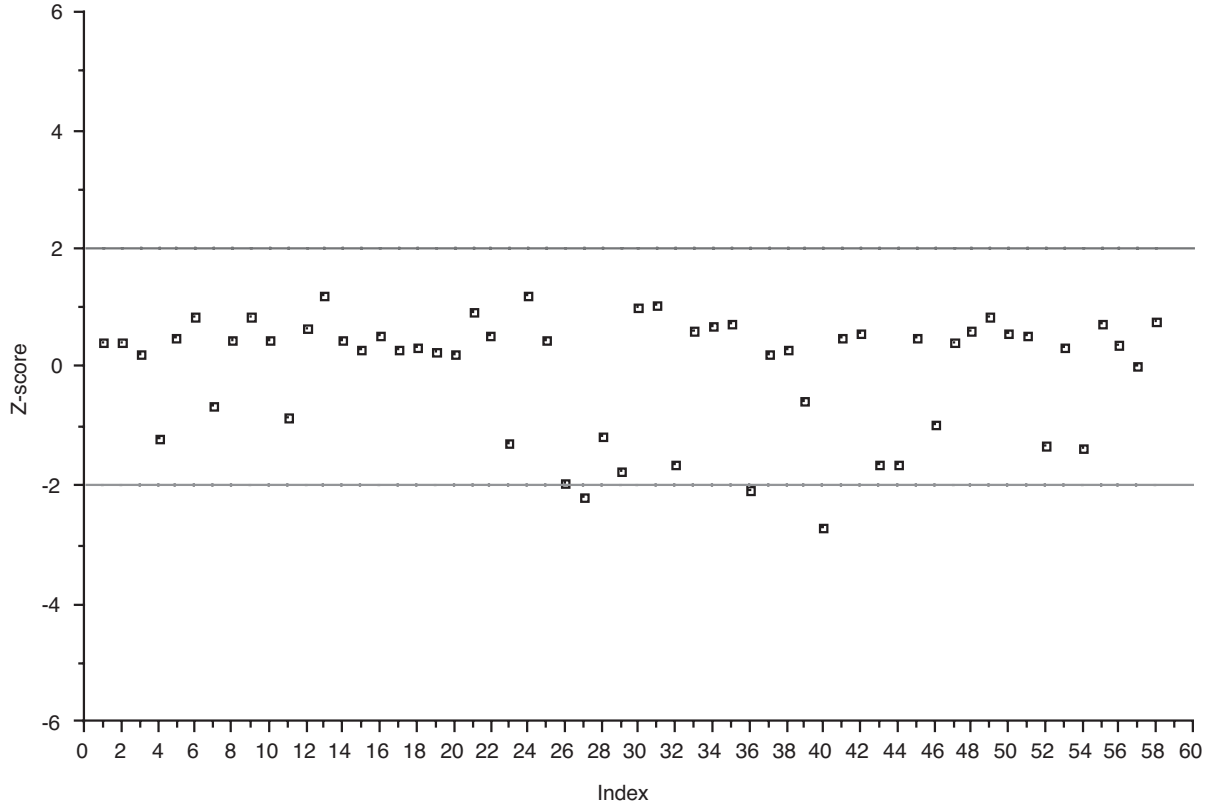


**Figure 4.** Quality control chart for student 6019 (INFIT MNSQ = 0.9, OUTFIT MNSQ = 1.0).

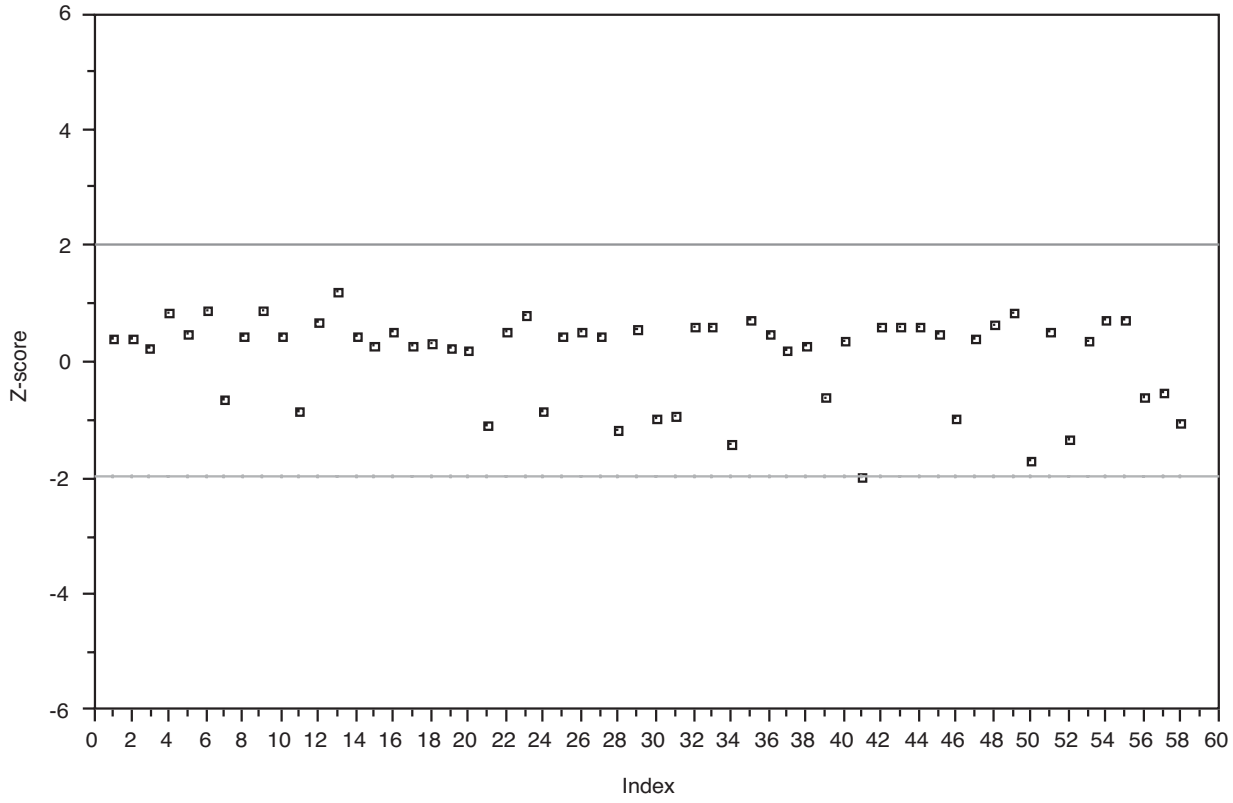


**Figure 5.** Quality control chart for student 1851 (INFIT MNSQ = 0.7, OUTFIT MNSQ = 0.6).

faculty consultant should give this student such a high rating on this question. (It should be noted that the figures use the index as the *x*-axis. See the tables for the correspondence between the index and actual question number.) Index is identical to question number when data for the student is not missing. (The AP scoring process ignores missing responses.) According to the student's self-reported demographic data, the student is a white male (Ethnic = 7, Gender = 1) whose best language is a language other than English (EBL = 3). For Student 2508, the fit statistics suggest that the multiple-choice and free-response questions may be providing inconsistent information or conflicting evidence regarding the student's level of English achievement. In cases like this, those in charge of monitoring quality control over the assessment system might want to review the student's performance before reporting an AP grade. The purpose would be to examine the student's pattern of performance across all 58 questions to determine whether the student's AP grade provides a valid indicator of the student's overall level of achievement and should be left to stand as it is, or perhaps whether one or more scores on individual questions are aberrant and require further investigation.

## Questions

Table 9 presents results from the calibration of the 58 questions included on the 1999 AP ELC assessment. The range of question difficulties is from –2.14 logits (Question 37) to 2.08 logits (Question 39), with 50 percent of the questions having difficulties between -0.54 and 0.75 logits. The mean difficulty of the questions (multiple-choice and free-responses combined) was set at zero,[8] and the average standard deviation is 0.97 logits. As shown in the variable map (Figure 2), the three free-response questions tend to be more difficult on average ($M$ = 1.07). The overall difference in difficulty between the questions is statistically significant, $\chi^2$ (57, $N$ = 58) = 72,684.9, $p < .01$, and the reliability of separation for the question difficulties is very high ($R$ = .99), indicating that the question difficulty

[8] When running FACETS analyses, it is customary to center all facets except one to establish a common origin, usually zero. If more than one facet is noncentered, then ambiguity may result since the frame of reference is not sufficiently constrained (Linacre, 1998). In this study, we centered questions at 0 logits and anchored faculty consultants at 0 logits, leaving the students facet noncentered. Consequently, the average question difficulty is 0, as is the average faculty consultant severity. Manipulating the center of the distribution does not substantively alter the results of this research; it simply produces an additive shift in the entire distribution of parameter estimates for a given facet.

TABLE 9

**Calibration of the Questions (1–55: Multiple-Choice, 56–58: Free-Response)**

| *Question* | *Difficulty Measure* | *SEM* | *INFIT MNSQ* | *OUTFIT MNSQ* |
|---|---|---|---|---|
| 1 | -.81 | .03 | 1.0 | .9 |
| 2 | -.73 | .03 | 1.0 | 1.0 |
| 3 | -1.89 | .04 | 1.0 | 1.0 |
| 4 | .67 | .02 | 1.0 | 1.0 |
| 5 | -.35 | .03 | 1.0 | 1.0 |
| 6 | .76 | .02 | 1.1 | 1.1 |
| 7 | 1.92 | .03 | 1.0 | 1.1 |
| 8 | -.49 | .03 | .9 | .8 |
| 9 | .77 | .02 | 1.1 | 1.1 |
| 10 | -.55 | .03 | .9 | .9 |
| 11 | 1.39 | .02 | 1.2 | 1.2 |
| 12 | .25 | .02 | 1.0 | 1.0 |
| 13 | 1.39 | .02 | 1.0 | 1.0 |
| 14 | -.54 | .03 | .9 | .9 |
| 15 | -1.35 | .04 | 1.0 | .9 |
| 16 | -.25 | .03 | .9 | .9 |
| 17 | -1.40 | .04 | 1.0 | .9 |
| 18 | -1.16 | .03 | .9 | .8 |
| 19 | -1.66 | .04 | .9 | .8 |
| 20 | -2.07 | .05 | .9 | .7 |
| 21 | .91 | .02 | 1.1 | 1.1 |
| 22 | -.25 | .03 | 1.0 | 1.0 |
| 23 | .57 | .02 | 1.1 | 1.1 |
| 24 | 1.42 | .02 | 1.0 | 1.0 |
| 25 | -.54 | .03 | .9 | .8 |
| 26 | -.29 | .03 | .9 | .8 |
| 27 | -.53 | .03 | 1.0 | .9 |
| 28 | .75 | .02 | .9 | .9 |
| 29 | -.10 | .03 | .9 | .9 |
| 30 | 1.08 | .03 | 1.0 | 1.0 |
| 31 | 1.16 | .02 | 1.0 | 1.0 |
| 32 | .06 | .03 | 1.0 | .9 |
| 33 | .02 | .03 | 1.0 | 1.0 |
| 34 | .36 | .02 | 1.0 | 1.0 |
| 35 | .45 | .02 | .9 | .9 |
| 36 | -.43 | .03 | 1.1 | 1.2 |
| 37 | -2.14 | .05 | 1.0 | .8 |
| 38 | -1.36 | .04 | 1.0 | 1.1 |
| 39 | 2.08 | .03 | 1.1 | 1.2 |
| 40 | -.96 | .03 | .9 | .9 |
| 41 | -.34 | .03 | 1.0 | 1.0 |
| 42 | -.01 | .03 | 1.0 | 1.0 |
| 43 | .04 | .03 | 1.0 | .9 |
| 44 | .06 | .03 | 1.0 | .9 |
| 45 | -.37 | .03 | 1.0 | 1.1 |
| 46 | 1.12 | .02 | 1.0 | 1.0 |
| 47 | -.72 | .03 | 1.0 | 1.1 |
| 48 | .12 | .03 | .9 | .9 |
| 49 | .73 | .02 | 1.0 | 1.0 |
| 50 | -.03 | .03 | 1.0 | 1.0 |
| 51 | -.22 | .03 | 1.0 | 1.0 |
| 52 | .45 | .03 | 1.1 | 1.1 |
| 53 | -1.07 | .04 | 1.0 | .9 |
| 54 | .43 | .03 | 1.1 | 1.1 |
| 55 | .39 | .03 | .9 | .9 |
| 56 | 1.17 | .01 | 1.1 | 1.1 |
| 57 | 1.15 | .01 | 1.1 | 1.1 |
| 58 | .89 | .01 | 1.2 | 1.2 |
| **Mean** | .00 | .03 | 1.0 | 1.0 |
| **SD** | .97 | .01 | .1 | .1 |

measures are very precisely estimated for this sample of 8,642 students (see Table 4).

The INFIT ($M$ = 1.0, $SD$ = 0.1) and OUTFIT mean-square summary statistics ($M$ = 1.0, $SD$ = 0.1) for the questions (shown at the bottom of Table 9) suggest that overall the fit of the data to the model is very good. Additionally, the fit mean-square statistics for each question were within acceptable quality-control limits of 0.8 to 1.2 with one exception: Question 20 (INFIT MNSQ = 0.9 and OUTFIT MNSQ = 0.7). Question 20 is somewhat overfitting, suggesting that this particular question shows some lack of independence in comparison with the other questions.[9] (This question was one of the easiest on the exam, which suggests that the problem with this question may be one of ineffective examinee targeting rather than dependency [J.M. Linacre, personal communication, March 14, 2002].) These findings suggest that, with the exception of Question 20, none of the questions appear to function in a redundant fashion. Furthermore, since none of the fit mean-square statistics is greater than 1.2, there appears to be little evidence of multidimensionality in this data. The 58 questions seem to work together to define the English achievement construct.[10] Overall, the scores on the multiple-choice questions correspond well to the ratings on the free-response questions. Therefore, it appears that scores on the multiple-choice questions can be meaningfully combined with the ratings on the free-response questions to produce a single summary measure (i.e., the AP grade) that can appropriately capture the essence of student performance across the 58 questions.

[9] As McNamara (1996) explained, overfitting items "are redundant items; they give us no information that the other items do not give; the pattern of response to these items is too predictable from the overall pattern of response to other items. Worse, they may signal items which have a dependency on other items built into them; for a example, if you can only get item 7 correct if you get item 6 correct, because understanding the point of item 7 depends on your having first understood the answer to item 6, then there will be too little variability in responses to item 7; the variability is constrained by the response to the previous item. Item 7 is not making an independent contribution to the measurement trait being measured by the test, and may therefore need to be revised or removed" (p. 176).

[10] As shown in Table 9, the three free-response questions (Q. 56–58) have INFIT MNSQ's of 1.1 or 1.2. According to Linacre (personal communication, March 14, 2002), these results suggest a slight misalignment between the multiple-choice questions and the free-response questions, which may be due to response format. One logit on the free-response questions would be equivalent to about 1.1 logits on the multiple-choice questions, so that combining the questions to calibrate them may exaggerate the misfit of the free-response questions.

## Faculty Consultants

The variable map in Figure 2 shows the severity measures for the 605 faculty consultants included in the study. Each star represents 18 faculty consultants. When we ran the FACETS analyses, we centered the faculty consultants at 0.00 logits. The faculty consultant severity measures range from -0.76 to 0.87 logits ($M$ = 0.00, $SD$ = 0.26), with 50 percent of the faculty consultants having severity measures between -0.17 and 0.17 logits. The separation index for faculty consultants is 1.57, indicating that within this sample of 605 faculty consultants there are about one-and-a-half statistically distinct strata of severity (see Table 4). The overall difference in severity between faculty consultants is statistically significant, $\chi^2$ (605, $N$ = 604) = 2,203.7, p < .01, and the reliability of separation is noticeably above what is desired ($R$ = .71). (The most desirable result would be to have a reliability of separation of 0.00, which would connote that the faculty consultants were interchangeable.) A reliability of separation of .71 means that, on average, there are discernible statistically significant differences between the severe and lenient faculty consultants.

Table 4 presents the summary statistics for the calibration of faculty consultants across all three free-response questions, and the calibrations of the faculty consultants within each free-response question. (As pointed out earlier, the faculty consultants were nested within the three free-response questions.) The average severities were not anchored to have a severity of 0.00 logits within each free-response question; this was not required because the faculty consultants are linked and calibrated through the multiple-choice questions. As is shown in Figure 2, the variability in faculty consultant severity is significantly smaller than the variability in question difficulties and student achievement levels. The faculty consultants were somewhat more lenient when rating essays for Question 58 ($M$ = -0.14, $SD$ = 0.26) than when rating essays for Question 56 ($M$ = 0.08, $SD$ = 0.27) and Question 57 ($M$ = 0.06, $SD$ = 0.24).

It is beyond the scope of this report to present the findings from a detailed analysis of faculty consultant fit. However, it is important to illustrate how persons in charge of monitoring quality control for an assessment system could use the fit information from a FACETS analysis to help them identify faculty consultants who are having trouble employing the scoring guidelines appropriately. Quality control tables (Tables 10, 11, and 12), as well as quality control charts (Figures 6, 7, and 8), have been constructed to illustrate the type of information that is available from FACETS analyses of score residuals. Quality control monitoring from a FACETS perspective involves examining the ratings a faculty consultant has assigned to determine to what extent the faculty consultant is using the AP scoring guidelines consistently across essays. By comparing the faculty consultant's observed and expected ratings for each essay rated, one can identify those faculty consultants who seem to be employing the scoring guidelines in an idiosyncratic way (i.e., unlike other faculty consultants). Additionally, by reviewing the output from an analysis of score residuals, one can pinpoint the particular essays that a faculty consultant rated inconsistently. For example, Faculty Consultant 347 has a noisy rating pattern (INFIT MNSQ = 3.0, OUTFIT MNSQ = 3.0). Generally, faculty consultants having infit mean-square statistics greater than 1 show more variation than expected in their ratings. They have assigned one or more ratings that are highly unexpected or surprising, given their level of severity and the difficulty of the free-response question they are scoring. If a faculty consultant has an outfit mean-square statistic that is larger than the infit mean-square statistic, then the faculty consultant will often have given a small number of highly unexpected or surprising ratings in the outermost categories of the scale. For the most part, the faculty consultant is internally consistent and uses the AP scoring guidelines appropriately, but occasionally the faculty consultant will give a rating that seems out of character with that consultant's other ratings. In some cases, faculty consultant misfit may also be attributable to uneven student performance across questions.

Table 10 presents a quality control chart inventorying the ratings that Faculty Consultant 347 gave and the standardized residual ($z$-score) associated with each rating. A $z$-score greater than 2.0 indicates an unexpectedly high rating for a student, given the faculty consultant's overall level of severity and the difficulty of the question, while a $z$-score less than -2.0 indicates an unexpectedly low rating for a student. Faculty Consultant 347 gave unexpectedly high ratings to five students' essays (739, 1561, 3047, 4015, and 5975) and unexpectedly low ratings to four students' essays (870, 3254, 3888, and 5396). In these types of cases, it is not feasible to adjust students' composite scores statistically to control for faculty consultant severity/leniency differences, since the faculty consultant does not show evidence of a consistent pattern of rating too severely or too leniently. Table 11 presents data for Faculty Consultant 370 (INFIT MNSQ = 1.0, OUTFIT MNSQ = 1.0) who has rated in a consistent fashion based on the MFRM model, while Table 12 provides data for Faculty Consultant 605 (INFIT MNSQ = 0.6, OUTFIT MNSQ = 0.6)

TABLE 10

**Quality Control Table for Faculty Consultant 347 (INFIT MNSQ = 3.0, OUTFIT MNSQ = 3.0)**

| *Index* | *Student ID* | *Student Gender* | *Student Race/Ethnicity* | *Student Best Language* | *Question* | *Faculty Consultant* | *Observed Rating* | *Expected Rating* | *Residual* | *Z-score* |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 473 | 1 | 4 | 1 | 57 | 347 | 6 | 5.66 | 0.34 | 0.28 |
| 2 | 631 | 2 | 7 | 1 | 57 | 347 | 4 | 2.25 | 1.75 | 1.82 |
| 3 | 739 | 2 | 7 | 1 | 57 | 347 | 7 | 4.01 | 2.99 | 2.66 |
| 4 | 870 | 2 | 7 | 1 | 57 | 347 | 2 | 4.48 | -2.48 | -2.15 |
| 5 | 1240 | 2 | 7 | 1 | 57 | 347 | 2 | 3.87 | -1.87 | -1.67 |
| 6 | 1255 | 2 | 7 | 1 | 57 | 347 | 2 | 4.11 | -2.11 | -1.86 |
| 7 | 1429 | 2 | 7 | 1 | 57 | 347 | 7 | 5.19 | 1.81 | 1.52 |
| 8 | 1529 | 2 | 7 | 1 | 57 | 347 | 1 | 3.10 | -2.10 | -1.95 |
| 9 | 1561 | 2 | 4 | 2 | 57 | 347 | 7 | 2.78 | 4.22 | 4.03 |
| 10 | 1925 | 2 | 7 | 1 | 57 | 347 | 4 | 5.44 | -1.44 | -1.20 |
| 11 | 3047 | 2 | 3 | 1 | 57 | 347 | 6 | 3.51 | 2.49 | 2.26 |
| 12 | 3056 | 2 | 4 | 2 | 57 | 347 | 4 | 4.08 | -0.08 | -0.07 |
| 13 | 3254 | 2 | 7 | 1 | 57 | 347 | 2 | 4.55 | -2.55 | -2.20 |
| 14 | 3266 | 2 | 7 | 1 | 57 | 347 | 3 | 3.55 | -0.55 | -0.50 |
| 15 | 3267 | 2 | 7 | 1 | 57 | 347 | 4 | 3.82 | 0.18 | 0.16 |
| 16 | 3664 | 1 | 4 | 1 | 57 | 347 | 3 | 4.45 | -1.45 | -1.26 |
| 17 | 3888 | 1 | 7 | 1 | 57 | 347 | 1 | 4.18 | -3.18 | -2.80 |
| 18 | 4015 | 2 | 7 | 1 | 57 | 347 | 8 | 4.71 | 3.29 | 2.81 |
| 19 | 4265 | 2 | 7 | 1 | 57 | 347 | 7 | 6.33 | 0.67 | 0.56 |
| 20 | 4845 | 2 | 7 | 1 | 57 | 347 | 5 | 3.91 | 1.09 | 0.97 |
| 21 | 5321 | 1 | 7 | 1 | 57 | 347 | 2 | 3.56 | -1.56 | -1.42 |
| 22 | 5396 | 2 | 4 | 1 | 57 | 347 | 1 | 4.34 | -3.34 | -2.92 |
| 23 | 5547 | 2 | 7 | 1 | 57 | 347 | 3 | 4.16 | -1.16 | -1.02 |
| 24 | 5622 | 1 | 7 | 1 | 57 | 347 | 4 | 4.85 | -0.85 | -0.73 |
| 25 | 5975 | 2 | 4 | 1 | 57 | 347 | 6 | 3.67 | 2.33 | 2.10 |
| 26 | 6207 | 2 | 7 | 1 | 57 | 347 | 7 | 5.91 | 1.09 | 0.91 |
| 27 | 6416 | 2 | 7 | 1 | 57 | 347 | 6 | 4.61 | 1.39 | 1.19 |
| 28 | 7181 | 1 | 7 | 1 | 57 | 347 | 6 | 4.74 | 1.26 | 1.07 |
| 29 | 7283 | 1 | 4 | . | 57 | 347 | 2 | 2.69 | -0.69 | -0.67 |
| 30 | 7337 | 2 | 7 | 1 | 57 | 347 | 4 | 3.99 | 0.01 | 0.01 |
| 31 | 7497 | 2 | 7 | 1 | 57 | 347 | 2 | 3.55 | -1.55 | -1.41 |
| 32 | 7729 | 2 | 7 | 1 | 57 | 347 | 7 | 4.70 | 2.30 | 1.97 |
| 33 | 7992 | 1 | 7 | 1 | 57 | 347 | 3 | 3.23 | -0.23 | -0.21 |

who may have rated too consistently (i.e., exhibited restriction of range in his/her ratings by not using all the rating categories included in the scoring guidelines). The information regarding model-data fit can also be presented graphically. Figures 6, 7, and 8 present quality control charts for Faculty Consultants 347, 370, and 605 with error bands inserted at +2 and –2.[11]

Often those in charge of monitoring quality control for an assessment system can use the detailed information contained in a ratings inventory in very practical ways to initiate meaningful improvements in the assessment system. For example, by studying the inventory of ratings for a faculty consultant, one could determine whether that consultant is having difficulty differentiating between certain points on the AP ELC rating scale (e.g., Are the unexpected ratings most often in the middle of the scale? Are they at one end of the scale?). One might also determine whether there are certain types of students that are difficult for the misfitting faculty consultant to evaluate reliably (e.g., Do the students included in the faculty consultant's inventory of unexpected ratings share common background characteristics: Are the students in urban schools? Do the students share the same racial or ethnic background?). Having access to this type of information on each misfitting faculty consultant provides a useful basis for

[11]The score residuals associated with ratings outside the error bands are more than two standard deviations of the modeled unit normal distribution of standardized residuals away from their expected value of 0.

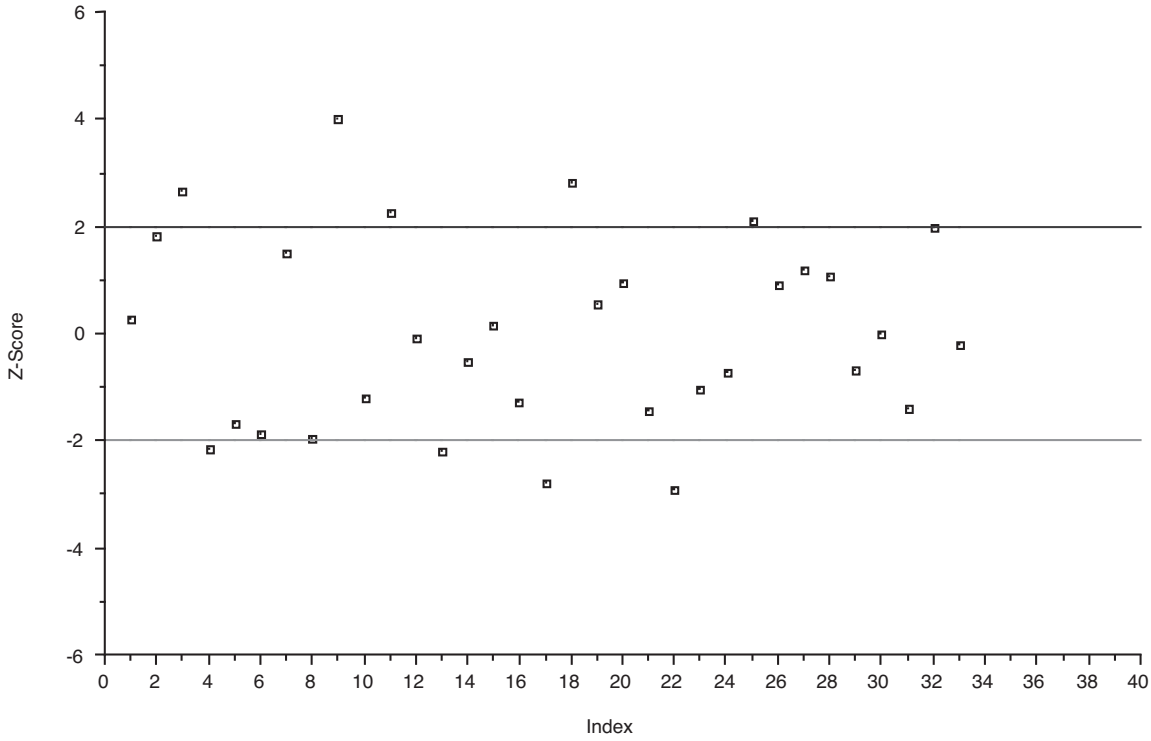


**Figure 6.** Quality control chart for faculty consultant 347 (INFIT MNSQ = 3.0, OUTFIT MNSQ = 3.0).

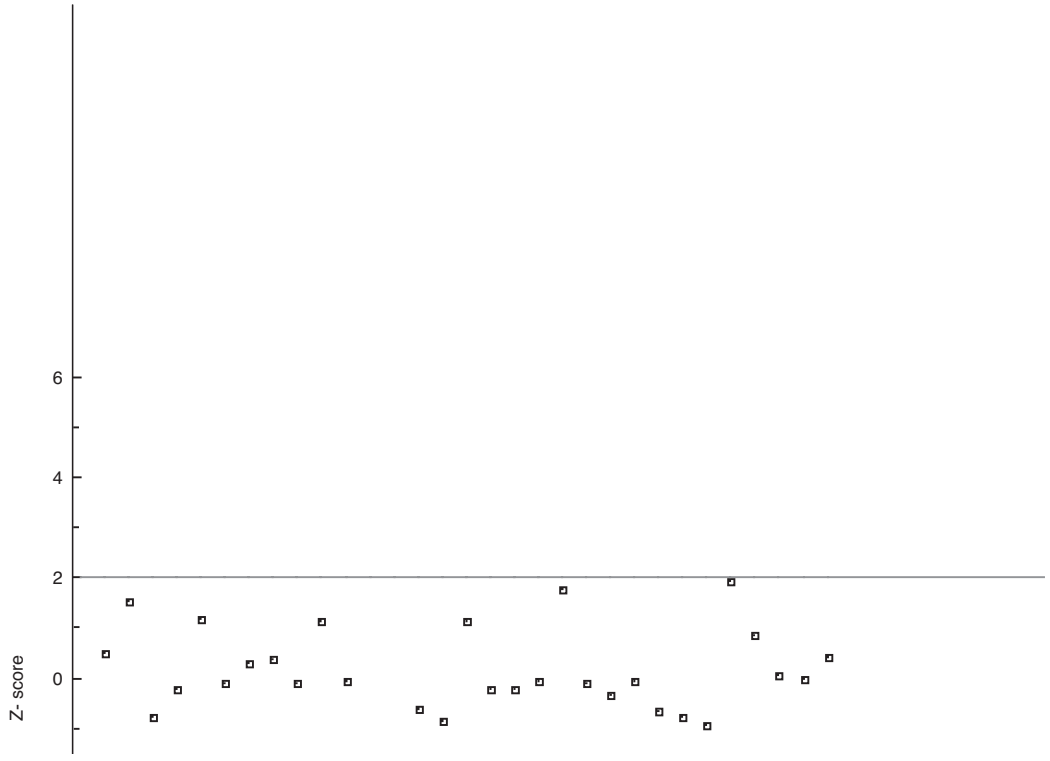


**Figure 7.** Quality control chart for faculty consultant 370 (INFIT MNSQ = 1.0, OUTFIT MNSQ = 1.0).

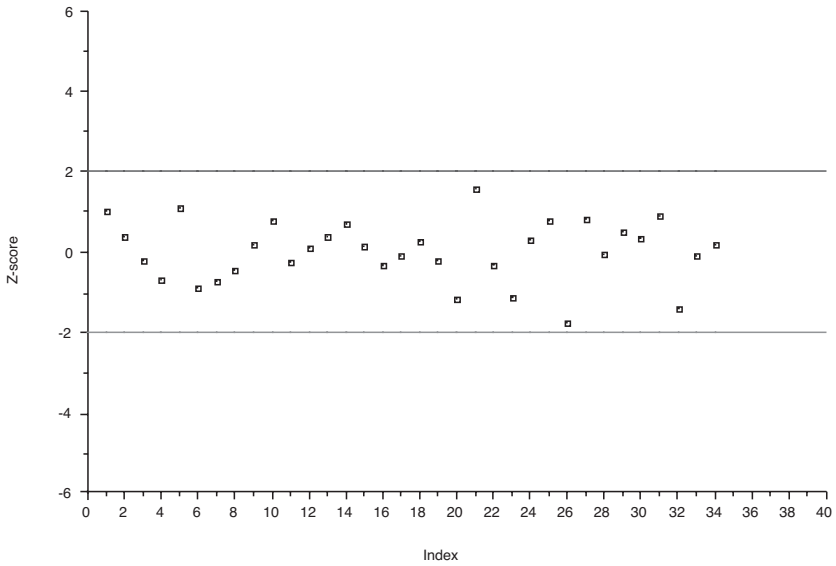


**Figure 8.** Quality control chart for faculty consultant 605 (INFIT MNSQ = 0.6, OUTFIT MNSQ = 0.6).

TABLE 11

**Quality Control Table for Faculty Consultant 370 (INFIT MNSQ = 1.0, OUTFIT MNSQ = 1.0)**

| *Index* | *Student ID* | *Student Gender* | *Student Race/Ethnicity* | *Student Best Language* | *Question* | *Faculty Consultant* | *Observed Rating* | *Expected Rating* | *Residual* | *Z-score* |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 64 | 2 | 7 | 1 | 58 | 370 | 4 | 3.44 | 0.56 | 0.51 |
| 2 | 175 | 2 | 4 | 1 | 58 | 370 | 7 | 5.15 | 1.85 | 1.55 |
| 3 | 375 | 1 | 6 | 1 | 58 | 370 | 4 | 4.89 | -0.89 | -0.76 |
| 4 | 761 | 2 | 7 | 1 | 58 | 370 | 4 | 4.26 | -0.26 | -0.22 |
| 5 | 1417 | 2 | 7 | 1 | 58 | 370 | 7 | 5.61 | 1.39 | 1.16 |
| 6 | 1436 | 2 | 7 | 1 | 58 | 370 | 8 | 8.08 | -0.08 | -0.09 |
| 7 | 1524 | 1 | 7 | 1 | 58 | 370 | 7 | 6.67 | 0.33 | 0.28 |
| 8 | 1967 | 2 | 8 | 1 | 58 | 370 | 7 | 6.56 | 0.44 | 0.37 |
| 9 | 2142 | 1 | 7 | 1 | 58 | 370 | 6 | 6.09 | -0.09 | -0.08 |
| 10 | 2291 | 1 | 7 | 1 | 58 | 370 | 6 | 4.66 | 1.34 | 1.15 |
| 11 | 3568 | 2 | . | 1 | 58 | 370 | 7 | 7.08 | -0.08 | -0.07 |
| 12 | 3582 | 2 | 7 | 1 | 58 | 370 | 4 | 6.50 | -2.50 | -2.10 |
| 13 | 3682 | 2 | 7 | 1 | 58 | 370 | 3 | 6.50 | -3.50 | -2.94 |
| 14 | 3693 | 2 | 7 | 1 | 58 | 370 | 6 | 6.71 | -0.71 | -0.60 |
| 15 | 3813 | 1 | . | 1 | 58 | 370 | 6 | 7.00 | -1.00 | -0.87 |
| 16 | 3917 | 1 | 7 | . | 58 | 370 | 6 | 4.68 | 1.32 | 1.14 |
| 17 | 3953 | 2 | 7 | 1 | 58 | 370 | 6 | 6.25 | -0.25 | -0.21 |
| 18 | 4237 | 2 | 7 | 1 | 58 | 370 | 6 | 6.24 | -0.24 | -0.20 |
| 19 | 4508 | 2 | 2 | 1 | 58 | 370 | 6 | 6.08 | -0.08 | -0.07 |
| 20 | 4509 | 1 | 2 | 1 | 58 | 370 | 6 | 4.01 | 1.99 | 1.77 |
| 21 | 4869 | 2 | 4 | 2 | 58 | 370 | 6 | 6.10 | -0.10 | -0.08 |
| 22 | 5112 | 1 | 7 | 1 | 58 | 370 | 6 | 6.38 | -0.38 | -0.32 |
| 23 | 5483 | 1 | 4 | 1 | 58 | 370 | 7 | 7.06 | -0.06 | -0.05 |
| 24 | 5949 | 1 | 7 | 1 | 58 | 370 | 5 | 5.77 | -0.77 | -0.64 |
| 25 | 6593 | 2 | 7 | 1 | 58 | 370 | 5 | 5.95 | -0.95 | -0.79 |
| 26 | 6758 | 2 | 2 | 1 | 58 | 370 | 5 | 6.14 | -1.14 | -0.95 |
| 27 | 6858 | 2 | 7 | 1 | 58 | 370 | 8 | 5.69 | 2.31 | 1.92 |
| 28 | 7057 | 2 | . | 2 | 58 | 370 | 7 | 6.00 | 1.00 | 0.84 |
| 29 | 7246 | 1 | 7 | 1 | 58 | 370 | 7 | 6.91 | 0.09 | 0.07 |
| 30 | 7689 | 2 | . | 1 | 58 | 370 | 7 | 7.02 | -0.02 | -0.02 |
| 31 | 7920 | 1 | 3 | 1 | 58 | 370 | 6 | 5.50 | 0.50 | 0.41 |

TABLE 12

**Quality Control Table for Faculty Consultant 605 (INFIT MNSQ = 0.6, OUTFIT MNSQ = 0.6)**

| *Index* | *Student ID* | *Student Gender* | *Student Race/Ethnicity* | *Student Best Language* | *Question* | *Faculty Consultant* | *Observed Rating* | *Expected Rating* | *Residual* | *Z-score* |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 9 | 2 | 7 | 1 | 58 | 605 | 5 | 3.85 | 1.15 | 1.03 |
| 2 | 594 | 2 | 6 | 1 | 58 | 605 | 5 | 4.56 | 0.44 | 0.38 |
| 3 | 595 | 2 | 7 | 1 | 58 | 605 | 4 | 4.22 | -0.22 | -0.20 |
| 4 | 1340 | 2 | 7 | 1 | 58 | 605 | 6 | 6.82 | -0.82 | -0.70 |
| 5 | 1351 | 2 | 7 | 1 | 58 | 605 | 7 | 5.66 | 1.34 | 1.11 |
| 6 | 1410 | 2 | 7 | 1 | 58 | 605 | 6 | 7.05 | -1.05 | -0.91 |
| 7 | 2277 | 2 | 7 | 1 | 58 | 605 | 5 | 5.88 | -0.88 | -0.73 |
| 8 | 2470 | 2 | . | 1 | 58 | 605 | 5 | 5.57 | -0.57 | -0.47 |
| 9 | 2551 | 2 | 7 | 1 | 58 | 605 | 5 | 4.78 | 0.22 | 0.19 |
| 10 | 2631 | 2 | 7 | 1 | 58 | 605 | 6 | 5.10 | 0.90 | 0.76 |
| 11 | 2671 | 1 | 7 | 1 | 58 | 605 | 6 | 6.31 | -0.31 | -0.26 |
| 12 | 2815 | 2 | 7 | 1 | 58 | 605 | 5 | 4.87 | 0.13 | 0.11 |
| 13 | 3002 | 1 | 7 | 1 | 58 | 605 | 4 | 3.58 | 0.42 | 0.38 |
| 14 | 3003 | 2 | 7 | 1 | 58 | 605 | 4 | 3.26 | 0.74 | 0.68 |
| 15 | 3016 | 2 | 7 | 1 | 58 | 605 | 5 | 4.85 | 0.15 | 0.13 |
| 16 | 3238 | 1 | 7 | 1 | 58 | 605 | 2 | 2.31 | -0.31 | -0.32 |
| 17 | 3514 | 1 | 4 | 1 | 58 | 605 | 5 | 5.11 | -0.11 | -0.09 |
| 18 | 3524 | 1 | 7 | 1 | 58 | 605 | 5 | 4.70 | 0.30 | 0.26 |
| 19 | 3559 | 2 | 7 | 1 | 58 | 605 | 5 | 5.27 | -0.27 | -0.22 |
| 20 | 3660 | 2 | 7 | 1 | 58 | 605 | 4 | 5.39 | -1.39 | -1.16 |
| 21 | 4019 | 1 | 7 | 1 | 58 | 605 | 8 | 6.10 | 1.90 | 1.58 |
| 22 | 4020 | 1 | 8 | 1 | 58 | 605 | 5 | 5.40 | -0.40 | -0.33 |
| 23 | 4269 | 2 | 7 | 1 | 58 | 605 | 4 | 5.37 | -1.37 | -1.14 |
| 24 | 4522 | 2 | 7 | 1 | 58 | 605 | 5 | 4.67 | 0.33 | 0.28 |
| 25 | 4744 | 2 | 4 | 1 | 58 | 605 | 5 | 4.13 | 0.87 | 0.77 |
| 26 | 4860 | 2 | 7 | 1 | 58 | 605 | 4 | 6.11 | -2.11 | -1.76 |
| 27 | 4916 | 2 | 7 | 1 | 58 | 605 | 6 | 5.05 | 0.95 | 0.80 |
| 28 | 5150 | 2 | 6 | 1 | 58 | 605 | 5 | 5.09 | -0.09 | -0.07 |
| 29 | 5338 | 2 | 7 | 1 | 58 | 605 | 8 | 7.48 | 0.52 | 0.48 |
| 30 | 5924 | 1 | . | 1 | 58 | 605 | 5 | 4.63 | 0.37 | 0.32 |
| 31 | 5994 | 2 | . | 1 | 58 | 605 | 9 | 8.27 | 0.73 | 0.89 |
| 32 | 6222 | 2 | 7 | 1 | 58 | 605 | 3 | 4.65 | -1.65 | -1.42 |
| 33 | 7055 | 2 | 4 | 2 | 58 | 605 | 4 | 4.10 | -0.10 | -0.09 |
| 34 | 8368 | 2 | 7 | 1 | 58 | 605 | 5 | 4.81 | 0.19 | 0.16 |

initiating targeted retraining, since it provides specific information to guide decision making regarding how best to work with a faculty consultant who is experiencing problems. Additionally, if it is possible to run FACETS analyses in "real time" (i.e., while the AP reading is taking place), then those in charge of monitoring quality control can use the faculty consultant fit information to identify early on those consultants who need additional training before they are allowed to score operationally.

## *Differential Facet Functioning Related to Faculty Consultants*

In this section of the report, we present findings from our investigation of the relationship between faculty consultant severity and student background characteristics. Specifically, we asked whether each faculty consultant maintained a uniform level of severity when rating essays of subgroups of students having different background characteristics, or whether some faculty

consultants appeared to exhibit differential severity/leniency, rating essays from some student subgroups more harshly or leniently than expected. To answer these questions, we included interaction effects in the measurement model. Specifically, facets were added to represent three background variables—student gender, student race/ethnicity, and student best language—so that bias analyses could be performed. Interaction effects provide evidence regarding differential faculty consultant functioning (DFCF) that corresponds conceptually to differential item functioning (DIF) analyses across relevant subgroups of students. It will become clear shortly that this process of exploring DFCF can be conceptualized as a method for studying residuals to determine whether there are identifiable patterns of unexpected ratings that are related to particular student subgroups.

### Differential Faculty Consultant Functioning Related to Student Gender

We conducted a faculty consultant crossed with student gender bias analysis to determine whether faculty consultants were rating essays composed by male and female students in a similar fashion, or whether some faculty consultants appeared to exhibit a bias toward (or against) essays composed by one or the other gender subgroup in the ratings they assigned. Specifically, we were interested in finding out whether any of the faculty consultants showed evidence of exercising differential severity/leniency, rating male students' essays (or female students' essays) more severely or leniently than expected, or whether each faculty consultant's level of severity/leniency was invariant across gender subgroups. Were there faculty consultants who were more prone to gender bias than other faculty consultants?

The first question we asked was whether, as a group, the faculty consultants showed a differential severity/leniency effect related to student gender. Table 13 provides summary statistics related to overall student gender differences in performance on the free-response questions included in the 1999 AP ELC Examination. These values represent the calibration values for the gender facet. FACETS computed an overall differential achievement measure for males based on the ratings that faculty consultants assigned all the essays that male students wrote for the three free-response questions. Similarly, FACETS computed an overall differential achievement measure for females based on the ratings that faculty consultants assigned all the essays that female students wrote for the three free-response questions. The differential achievement measure for males was 0.02 logits (*SE* = 0.01), while the differential achievement measure for females was –0.02 logits (*SE* = 0.01). Even though there is a statistically significant difference between these achievement measures ($\chi^2$ (1, $N$ = 2) = 53.5, $p$ < .01), the difference is so small on the logit scale that it does not warrant substantive interpretation. (Differences less than 0.30 logits are usually not substantively meaningful.) Overall, it does not appear that the faculty consultants showed a differential severity/leniency effect related to student gender.

In the process of carrying out a bias analysis, FACETS looks at all possible combinations of elements of facets included in the analysis that may have an unexpected effect on the estimation of student achievement. When investigating a potential differential severity/leniency effect related to student gender, the computer program would examine pairs of facets (i.e., each faculty consultant crossed with student gender combination) to pinpoint ratings that may show gender bias (i.e., ratings that show a consistent pattern related to student gender that is different from the pattern revealed in the overall analysis). Table 14 provides summary statistics for all of the subgroups. For gender, only 1.5 percent of the possible interaction effects (faculty consultant crossed with student gender) are statistically significant; the overall test that the interaction effects significantly vary from zero is also not statistically significant, $\chi^2$ (1210, $N$ = 1209) = 772.6, $p$ = ns. (There were 1210 interaction terms: 605 faculty consultants crossed with 2 student gender categories.) While the evidence strongly suggests that there is not a group-level differential severity/leniency effect related to student gender, we wanted to know whether there were individual faculty consultants that may have displayed differential severity/leniency in their ratings. The FACETS analysis identified 18 faculty consultants (out of 605) that, based on statistical criteria, may have exhibited differential

TABLE 13

**Differences in Faculty Consultants' Ratings Related to Student Gender**

| *Student Gender* | *Measure (SEM)* | *Mean Differences* | |
|---|---|---|---|
| | | *Male* | *Female* |
| Male | .02 (.01) | X | .04 |
| Female | -.02 (.01) | | X |
| Chi-Square | 53.5* | | |
| df | 1 | | |

* p < .01

TABLE 14

**Summary of Differential Faculty Consultant Functioning (Interaction Terms) by Student Subgroups**

| | *Student Gender* | *Student Race/Ethnicity* | *Student Best Language (EBL)* |
|---|---|---|---|
| Count of Interaction Terms | 1210 | 3176 | 413 |
| \| Z \| > = 2.0 | 18 | 118 | 3 |
| % statistically significant | 1.5% | 3.72% | 0.73% |
| Chi-Square | 772.6 | 2937.2 | 197.2 |
| df | 1209 | 3175 | 412 |

severity/leniency effects related to student gender. Table 15 provides summary DFCF statistics for selected faculty consultants from the pool of 18 faculty consultants with significant (|Z| > = 2.0) interaction terms.

By examining interaction effects, one can identify patterns based on subgroup membership that are related to the discrepancies between observed and expected ratings produced by the many-faceted Rasch model. In order to illustrate this process, a quality control table was constructed for Faculty Consultant 108 (Table 16). This information is presented graphically in Figure 9. Faculty Consultant 108 provides an interesting case study for exploring DFCF related to student gender. Based on the overall fit statistics (INFIT MNSQ = 1.1, OUTFIT MNSQ = 1.1), Faculty Consultant 108 does not appear to be rating in an unusual fashion. However, when the interaction between faculty consultant and student gender is explicitly estimated and examined, a slightly different story emerges. The *z*-statistics provide a test of the statistical significance of the interactions and suggest that Faculty Consultant 108 tended to rate the male students' essays higher than expected (*z*-statistic, males = -2.00); based on the model, we would predict that Faculty Consultant 108 would have an expected mean rating of 4.56—instead, the observed mean was 5.33 (0.77 points higher). Turning to the female students' essays, the difference of -0.30 between the observed mean (4.83) and expected mean (5.13) is not statistically significant (*z*-statistic, females = 1.25). An examination of Figure 9 clearly shows that Faculty Consultant 108 tended to consistently assign higher-than-expected ratings to the nine male students' essays. Figure 9 also shows that Faculty Consultant 108 was not as consistent in assigning lower-than-expected ratings to the 23 female students' essays.

Figure 9 highlights the importance of exploring not only mean differences between observed and expected ratings within each subgroup category but also the variability and spread of residuals within subgroups. Ultimately, differential faculty consultant functioning involves looking at discrepancies between observed and expected ratings at the individual level. As pointed out many years ago by Wright (1984, p. 285),

> …bias found for groups is never uniformly present among members of the groups or uniformly absent among those not in the group. For the analysis of item bias to do individuals any good, say, by removing the bias from their measures, it will have to be done on the individual level.

This point also applies to rater-mediated assessments and potential faculty consultant bias. It is very important to conduct group-level analyses of DFCF, but the full interpretation of these effects requires a detailed examination of residuals for each faculty consultant. This also highlights the importance of using caution if

TABLE 15

**Summary of Differential Faculty Consultant Functioning Statistics (Interactions) for Selected Faculty Consultants by Student Gender**

| *Faculty Consultant* | *INFIT MNSQ* | *OUTFIT MNSQ* | *Student Gender Subgroup* | *Count* | *Mean Observed* | *Mean Expected* | *Mean Residual* | *Bias Logit* | *SE* | *Z-statistic* |
|---|---|---|---|---|---|---|---|---|---|---|
| 108 | 1.1 | 1.1 | Male | 9 | 5.33 | 4.56 | .77 | -.56 | .28 | -2.00* |
| | | | Female | 23 | 4.83 | 5.13 | -.30 | .23 | .18 | 1.25 |
| 142 | 2.0 | 1.9 | Male | 16 | 4.44 | 5.07 | -.63 | .49 | .22 | 2.19* |
| | | | Female | 24 | 5.75 | 5.33 | .42 | -.31 | .18 | -1.77 |
| 146 | 2.1 | 2.1 | Male | 13 | 4.54 | 5.19 | -.65 | .48 | .24 | 2.01* |
| | | | Female | 28 | 5.43 | 5.12 | .31 | -.22 | .16 | -1.38 |

* | Z | > = 2.00

TABLE 16

**Quality Control Table for Faculty Consultant 108 (INFIT MNSQ = 1.1, OUTFIT MNSQ = 1.1) (DFCF interaction *z*-statistics: males = -2.00, females = 1.25)**

| *Index* | *Student ID* | *Student Gender* | *Student Ethnicity* | *Student EBL* | *Question* | *Faculty Consultant* | *Observed Rating* | *Expected Rating* | *Residual* | *Z-score* |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 375 | 1 | 6 | 1 | 57 | 108 | 5 | 3.86 | 1.14 | 1.02 |
| 2 | 2093 | 1 | 7 | 1 | 57 | 108 | 5 | 4.37 | 0.63 | 0.55 |
| 3 | 2128 | 1 | 7 | 1 | 57 | 108 | 5 | 3.72 | 1.28 | 1.16 |
| 4 | 3473 | 1 | 7 | 1 | 57 | 108 | 5 | 4.64 | 0.36 | 0.31 |
| 5 | 3670 | 1 | 7 | 1 | 57 | 108 | 7 | 4.99 | 2.01 | 1.70 |
| 6 | 3744 | 1 | 7 | 1 | 57 | 108 | 6 | 5.21 | 0.79 | 0.67 |
| 7 | 4109 | 1 | 7 | 1 | 57 | 108 | 6 | 4.76 | 1.24 | 1.06 |
| 8 | 5147 | 1 | . | 2 | 57 | 108 | 6 | 5.75 | 0.25 | 0.21 |
| 9 | 7663 | 1 | 7 | 1 | 57 | 108 | 3 | 3.75 | -0.75 | -0.67 |
| 10 | 351 | 2 | 7 | 1 | 57 | 108 | 3 | 4.48 | -1.48 | -1.29 |
| 11 | 459 | 2 | 2 | 1 | 57 | 108 | 4 | 5.67 | -1.67 | -1.39 |
| 12 | 1200 | 2 | 7 | 1 | 57 | 108 | 4 | 5.30 | -1.30 | -1.09 |
| 13 | 1613 | 2 | 7 | 1 | 57 | 108 | 4 | 4.18 | -0.18 | -0.16 |
| 14 | 1689 | 2 | 2 | 1 | 57 | 108 | 4 | 3.68 | 0.32 | 0.29 |
| 15 | 3228 | 2 | 7 | 1 | 57 | 108 | 2 | 3.42 | -1.42 | -1.30 |
| 16 | 3628 | 2 | 7 | 1 | 57 | 108 | 4 | 4.98 | -0.98 | -0.83 |
| 17 | 3658 | 2 | 7 | 1 | 57 | 108 | 5 | 5.18 | -0.18 | -0.15 |
| 18 | 4853 | 2 | 4 | 1 | 57 | 108 | 5 | 5.77 | -0.77 | -0.64 |
| 19 | 5040 | 2 | 4 | 1 | 57 | 108 | 6 | 5.93 | 0.07 | 0.06 |
| 20 | 5231 | 2 | 7 | 1 | 57 | 108 | 6 | 7.49 | -1.49 | -1.38 |
| 21 | 5298 | 2 | 8 | 1 | 57 | 108 | 9 | 7.62 | 1.38 | 1.31 |
| 22 | 5453 | 2 | . | 1 | 57 | 108 | 5 | 4.49 | 0.51 | 0.44 |
| 23 | 5471 | 2 | . | 1 | 57 | 108 | 6 | 5.90 | 0.10 | 0.08 |
| 24 | 5760 | 2 | 7 | 1 | 57 | 108 | 7 | 5.56 | 1.44 | 1.20 |
| 25 | 6140 | 2 | 7 | 1 | 57 | 108 | 7 | 4.00 | 3.00 | 2.67 |
| 26 | 6525 | 2 | 7 | 1 | 57 | 108 | 6 | 5.43 | 0.57 | 0.48 |
| 27 | 7325 | 2 | 7 | 1 | 57 | 108 | 6 | 5.09 | 0.91 | 0.77 |
| 28 | 7560 | 2 | 7 | 1 | 57 | 108 | 4 | 4.00 | 0.00 | 0.00 |
| 29 | 8057 | 2 | 7 | 1 | 57 | 108 | 3 | 5.40 | -2.40 | -2.01 |
| 30 | 8389 | 2 | . | 1 | 57 | 108 | 5 | 5.97 | -0.97 | -0.81 |
| 31 | 8404 | 2 | 2 | 2 | 57 | 108 | 1 | 2.91 | -1.91 | -1.80 |
| 32 | 8440 | 2 | 7 | 1 | 57 | 108 | 5 | 5.50 | -0.50 | -0.42 |

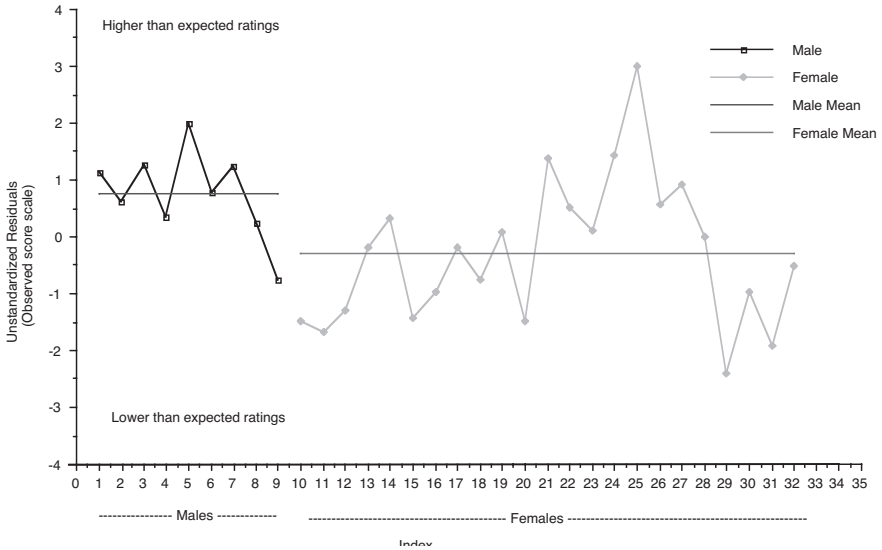


**Figure 9.** Faculty consultant 108 (DFCF, interaction *z*-statistics: males = -2.00, females = 1.25).

routine statistical adjustments are made for faculty consultant severity.

### Differential Faculty Consultant Functioning Related to Student Race/Ethnicity

We conducted a faculty consultant crossed with student race/ethnicity bias analysis to determine whether faculty consultants were rating essays composed by students from various racial/ethnic backgrounds in a similar fashion, or whether some faculty consultants appeared to exhibit a bias in their ratings toward (or against) essays composed by students from certain racial/ethnic subgroups. Specifically, we were interested in finding out whether any of the faculty consultants showed evidence of exercising differential severity/leniency, rating essays composed by students from a particular racial/ethnic background more severely or leniently than expected, or whether each faculty consultant's level of severity/leniency was invariant across race/ethnicity subgroups. Were there faculty consultants who were more prone to race/ethnicity bias than other faculty consultants?

First, we asked whether, as a group, the faculty consultants showed a differential severity/leniency effect related to student race/ethnicity. Table 17 provides summary statistics related to mean differences in the ratings assigned essays that were written by students in the eight race/ethnicity subgroups identified in the AP questionnaire for students. It should be stressed that these categories are based on self-report data, and that some students may not perceive the categories as being mutually exclusive. FACETS computed an overall differential achievement measure for each of the eight race/ethnicity subgroups based on the ratings that faculty consultants assigned all the essays that students in each race/ethnicity subgroup wrote for the three free-response questions. As was found with the earlier analyses of student gender differences, although there are statistically significant differences between student race/ethnicity subgroup measures, these differences tend to be very small on the logit scale. The largest mean differences in achievement were between (Ethnic = 1, American Indian or Alaska Native) and (Ethnic = 5, Puerto Rican), and between (Ethnic = 5, Puerto Rican) and (Ethnic = 6, South American, Latin American, Central American, or other Hispanic). In both cases, the differences between subgroup achievement measures were 0.28 logits. (Again, differences less than 0.30 logits are usually not substantively meaningful.) It does not appear that, overall, the faculty consultants showed a differential severity/leniency effect related to student race/ethnicity. As shown in Table 14, only 3.72 percent of the possible interaction effects (faculty consultant crossed with student race/ethnicity) are statistically significant; the overall test that the interaction effects significantly vary from zero is also not statistically significant, $\chi^2$ (3175, $N$ = 3176) = 2937.2, $p$ = ns.

While the evidence suggests that there is not a group-level differential severity/leniency effect related to student race/ethnicity, we wanted to know whether there were individual faculty consultants that may have displayed differential severity/leniency in their ratings. The FACETS analysis identified some faculty consultants who appeared to have exhibited differential severity/leniency effects related to student race/ethnicity. Table 18 provides summary DFCF statistics for two of the faculty consultants who had significant ($|Z| > = 2.0$) interaction terms. Detailed quality control tables and charts can be constructed for further study. (In order to save space, these tables and charts are not presented here.) It is important to emphasize that in many of these cases, the

TABLE 17

**Differences in Faculty Consultants' Ratings Related to Student Race/Ethnicity**

| | | *Mean Differences* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Student Race/Ethnicity* | *Measure (SEM)* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* |
| 1 | .12 (.08) | X | .17 | .14 | .00 | .28 | .12 | .16 | .08 |
| 2 | -.05 (.03) | | X | -.01 | -.17 | .11 | -.05 | -.01 | -.09 |
| 3 | -.04 (.03) | | | X | -.16 | .12 | -.04 | .00 | -.08 |
| 4 | .12 (.02) | | | | X | -.16 | .12 | .16 | .08 |
| 5 | -.16 (.07) | | | | | X | .28 | -.12 | -.20 |
| 6 | .00 (.04) | | | | | | X | .04 | -.04 |
| 7 | -.04 (.03) | | | | | | | X | .00 |
| 8 | .04 (.03) | | | | | | | | X |
| Chi-Square | 86.9* | | | | | | | | |
| df | 7 | | | | | | | | |

* p < .01

Note: Student race/ethnicity subgroup designations are as follows: 1 = American Indian or Alaska Native; 2 = Black or African American; 3 = Mexican American or Chicano; 4 = Asian, Asian American, or Pacific Islander; 5 = Puerto Rican; 6 = South American, Latin American, Central American, or other Hispanic; 7 = White; 8 = Other.

TABLE 18

**Summary of Differential Faculty Consultant Functioning Statistics (Interactions) for Selected Faculty Consultants by Student Race/Ethnicity**

| *Faculty Consultant* | *INFIT MNSQ* | *OUTFIT MNSQ* | *Student Race/ Ethnicity Subgroup* | *Count* | *Mean Observed* | *Mean Expected* | *Mean Residual* | *Bias Logit* | *SE* | *Z-statistics* |
|---|---|---|---|---|---|---|---|---|---|---|
| 236 | 2.0 | 1.9 | 1 | 1 | 2.00 | 4.40 | -2.43 | 2.17 | 1.12 | 1.94 |
| | | | 2 | 0 | X | X | X | X | X | X |
| | | | 3 | 2 | 5.00 | 3.95 | 1.05 | -.79 | .60 | -1.32 |
| | | | 4 | 7 | 6.14 | 5.03 | 1.12 | -.80 | .32 | -2.50* |
| | | | 5 | 0 | X | X | X | X | X | X |
| | | | 6 | 0 | X | X | X | X | X | X |
| | | | 7 | 28 | 4.79 | 4.83 | -.05 | .04 | .16 | .22 |
| | | | 8 | 1 | 5.00 | 5.80 | -.80 | .56 | .86 | .60 |
| 621 | 1.8 | 1.8 | 1 | 0 | X | X | X | X | X | X |
| | | | 2 | 1 | 4.00 | 6.00 | -1.96 | 1.43 | .89 | 1.60 |
| | | | 3 | 2 | 3.50 | 4.70 | -1.18 | .92 | .65 | 1.43 |
| | | | 4 | 4 | 3.00 | 4.70 | -1.70 | 1.40 | .49 | 2.88* |
| | | | 5 | 0 | X | X | X | X | X | X |
| | | | 6 | 1 | 2.00 | 3.60 | -1.62 | 1.54 | 1.12 | 1.37 |
| | | | 7 | 24 | 6.17 | 5.52 | .64 | -.46 | .17 | -2.66* |
| | | | 8 | 1 | 6.00 | 6.40 | -.39 | .27 | .83 | .33 |

* | Z | > = 2.00

Note: Student race/ethnicity subgroup designations are as follows: 1 = American Indian or Alaska Native; 2 = Black or African American; 3 = Mexican American or Chicano; 4 = Asian, Asian American, or Pacific Islander; 5 = Puerto Rican; 6 = South American, Latin American, Central American, or other Hispanic; 7 = White; 8 = Other.

faculty consultant rated a small number of essays composed by students whose selected a race/ethnicity category other than 7 (white) on their AP questionnaire.

Consequently, the designation of these faculty consultants as "biased" in their ratings of certain race/ethnicity subgroups is very preliminary and would need further verification by having the faculty consultant rate more essays written by students in these particular subgroups to obtain a larger, more representative sample of his or her rating behavior upon which to base judgments regarding possible bias. Nonetheless, we include below a discussion of the results from the bias analysis to illustrate how one could use the output from the analysis to gain an understanding of the nature of differential severity/leniency and its potential impact on the various student race/ethnicity subgroups. In interpreting these analyses it important to keep in mind that some of the categories within a subgroup have very small *N*s. Additional examination of the essays is required to determine whether or not these interaction effects are substantively significant.

Faculty Consultant 236 shows evidence of having given some unexpected or surprising ratings (INFIT MNSQ = 2.0, OUTFIT MNSQ = 1.9). The *z*-statistics suggest that this faculty consultant's ratings of the seven essays written by Asian, Asian American, and Pacific Islander students (Ethnic = 4) were higher than expected (Mean Observed = 6.14, Mean Expected = 5.03). The model predicted that the faculty consultant would assign ratings to these students' essays that were, on average, 1.11 points lower than those actually given. In the case of Faculty Consultant 621, the fit mean-square statistics suggest some degree of unexpectedness in the faculty consultant's ratings (INFIT MNSQ = 1.8, OUTFIT MNSQ = 1.8). This faculty consultant tended to assign lower-than-expected ratings to the four essays written by Asian, Asian American, and Pacific Islander students (Mean Observed = 3.00, Mean Expected = 4.70). The model predicted that the faculty consultant would assign ratings to these students' essays that were, on average, 1.70 points higher than those actually given. By contrast, Faculty Consultant 236 gave higher-than-expected ratings to the 24 essays written by White (Ethnic = 7) students (Mean Observed = 6.17, Mean Expected = 5.52). Based on modeled predictions, the faculty consultant tended to assign ratings to these students' essays that were, on average, 0.65 points higher than expected.

## Differential Faculty Consultant Functioning Related to Student Best Language

We conducted a faculty consultant x student best language bias analysis to determine whether faculty consultants were rating essays composed by students from different language backgrounds in a similar fashion, or whether some faculty consultants appeared to exhibit a bias in their ratings toward (or

against) essays composed by students whose first language may not be English. Specifically, we were interested in finding out whether any of the faculty consultants showed evidence of exercising differential severity/leniency, rating essays composed by students from a particular language background more severely or leniently than expected, or whether each faculty consultant's level of severity/leniency was invariant across language subgroups. Were there faculty consultants who were more prone to bias based on student language background than other faculty consultants?

We asked whether, as a group, the faculty consultants showed a differential severity/leniency effect related to student best language. Table 19 provides summary statistics related to mean differences in ratings assigned essays that were written by students in the three EBL groups identified in the AP questionnaire for students. (The three EBL groups were [1] English, [2] English and another language about the same, and [3] Another language.) It should be stressed that these EBL groupings are also based on self-report data, and that some students may not have responded accurately. FACETS computed an overall differential achievement measure for each of the three language subgroups based on the ratings that faculty consultants assigned all the essays that students in each language subgroup wrote for the three free-response questions. All of the contrasts of EBL = 1 (English) and EBL = 2 (English and another language about the same) with EBL = 3 (Another language) are statistically significant and larger than 0.30 logits. The EBL = 3 subgroup is quite small ($N$ = 36), but a comparison of the three language subgroup differential achievement measures suggests that students who place themselves in this category tend on average to receive higher ratings on their essays than the students in the other two language subgroups. As shown in Table 14, only 0.73 percent of the possible interaction effects (faculty consultant x student EBL) are statistically significant; the overall test that the interaction effects significantly vary from zero is also not statistically significant, $\chi^2$ (413, $N$ = 412) = 197.2, $p$ = ns.

While the evidence suggests that there is not a group-level differential severity/leniency effect related to student best language, we wanted to know whether there were individual faculty consultants who may have displayed differential severity/leniency in their ratings. The FACETS analysis identified three faculty consultants who appeared to have exhibited differential severity/leniency effects related to student best language. Table 20 provides summary DFCF statistics for the faculty consultants with significant (|Z| > = 2.0) interaction terms. Detailed quality control tables

TABLE 19

**Differences in Faculty Consultants' Ratings Related to Student Best Language**

| *Student Best Language (EBL)* | *Measure (SEM)* | *Mean Differences* | | |
|---|---|---|---|---|
| | | *1* | *2* | *3* |
| 1 | -.28 (.01) | X | -.12 | -.72 |
| 2 | -.16 (.04) | | X | -.60 |
| 3 | .44 (.14) | | | X |
| Chi-Square | 34.2* | | | |
| df | 2 | | | |

*p < .01.77

Note: Student best language (EBL) subgroup designations are as follows: 1 = English; 2 = English and another language about the same; 3 = Another language.

TABLE 20

**Summary of Differential Faculty Consultant Functioning Statistics (Interactions) for Selected Faculty Consultants by Student Best Language**

| *Faculty Consultant* | *INFIT MNSQ* | *OUTFIT MNSQ* | *Student Best Language (EBL) Subgroup* | *Count* | *Mean Observed* | *Mean Expected* | *Mean Residual* | *Bias Logit* | *SE* | *Z-statistics* |
|---|---|---|---|---|---|---|---|---|---|---|
| 705 | 1.0 | 1.0 | 1 | 33 | 4.36 | 4.53 | -.17 | .13 | .15 | .83 |
| | | | 2 | 5 | 5.00 | 3.90 | 1.09 | -.83 | .38 | -2.17* |
| | | | 3 | 0 | X | X | X | X | X | X |
| 325 | 1.3 | 1.3 | 1 | 48 | 4.69 | 4.46 | .23 | -.17 | .13 | -1.38 |
| | | | 2 | 6 | 2.83 | 4.08 | -1.25 | 1.09 | .40 | 2.69* |
| | | | 3 | 0 | X | X | X | X | X | X |
| 428 | 1.4 | 1.5 | 1 | 78 | 4.86 | 4.78 | .08 | -.06 | .10 | -.62 |
| | | | 2 | 2 | 3.50 | 5.65 | -2.14 | 1.62 | .65 | 2.50* |
| | | | 3 | 1 | 3.00 | 4.20 | -1.19 | .98 | .94 | 1.04 |

*| Z | > = 2.00

Note: Student best language (EBL) subgroup designations are as follows: 1 = English; 2 = English and another language about the same; 3 = Another language.

and charts can be constructed for further study. (In order to save space, these tables and charts are not presented here.) It is important to emphasize that in all three cases, each faculty consultant rated a small number of essays composed by students whose best language was English and another language, or another language other than English. Consequently, the designation of these faculty consultants as "biased" in their ratings is very preliminary and would need further verification by having the faculty consultant rate more essays written by students in these particular subgroups to obtain a larger, more representative sample of his or her rating behavior upon which to base judgments regarding possible bias. Nonetheless, we include below a discussion of the results from the bias analysis to illustrate how one could use the output from the analysis to gain an understanding of the nature of differential severity/leniency and its potential impact on the various student language subgroups.

Faculty Consultant 705 appears to have used the AP score guidelines in a consistent fashion (INFIT MNSQ = 1.0, OUTFIT MNSQ = 1.0). However, the *z*-statistics suggest that this faculty consultant's ratings of the five essays written by students who consider English and another language to be their "best language" (EBL = 2) were higher than expected (Mean Observed = 5.00, Mean Expected = 3.90). The model predicted that the faculty consultant would assign ratings to these students' essays that were, on average, 1.09 points lower than those actually given. Faculty Consultant 325 also appears to have used the AP score guidelines in a consistent manner overall (INFIT MNSQ = 1.3, OUTFIT MNSQ = 1.3). This faculty consultant tended to assign lower-than-expected ratings to the six essays written by students who consider English and another language to be their "best language" (EBL = 2) (Mean Observed = 2.83, Mean Expected = 4.08). The model predicted that the faculty consultant would assign ratings to these students' essays that were, on average, 1.25 points higher than those actually given. Similarly, Faculty Consultant 428 gave lower-than-expected ratings to the two essays written by students who consider English and another language to be their "best language" (EBL = 2) (Mean Observed = 3.50, Mean Expected = 5.65). Based on modeled predictions, the faculty consultant tended to assign ratings to these students' essays that were, on average, 2.15 points lower than expected.

## *Impact on AP Composite Scores and AP Grades of Adjusting for Differences in Faculty Consultant Severity*

This section focuses on the empirical analyses of the effects of differences in faculty consultant severity on the AP composite scores and AP grades. Estimates of student achievement obtained from the FACETS analyses reported in the previous section of this report were used to provide theta values that were adjusted for faculty consultant severity effects. Equations 8 and 9 were employed to obtain these adjusted ratings. We used these theta values to obtain AP composite scores and AP grades that were adjusted for differences in faculty consultant severity. These adjusted composite scores and grades were then compared to the actual AP composite scores and grades that students obtained. The use of the faculty consultant adjusted thetas to produce values on the 9-point AP score scale is a straight-forward generalization of formula scoring proposed by Lord (1980), and described in Crocker and Algina (1986) for the case of dichotomous items. It should be pointed out that the multiple-choice questions were not rescored; rather, the original weighted values (Weighted Section I) were used to compute the actual AP composite scores and grades as well as the adjusted AP composite scores and grades. Only the three ratings on the essays for the three free-response questions were adjusted for faculty consultant severity using each student's theta value.

The scoring process used for the AP ELC Exam is shown in Figure A1 (see Appendix). The multiple-choice section is based on a correction process. This correction is called the "rights minus wrongs correction" (Crocker and Algina, 1986, p. 400). The formula for calculating this is:

$$X_c = R - W/(k\text{-}1)$$

where $X_c$ is the score corrected for guessing, R is the number of correct answers, W is the number of incorrect answers, and $k$ is the number of alternatives per question. The AP scoring process ignores missing responses. The formula score is then multiplied by 1.2272 in order to obtained the Weighted Section I Score. As shown in Figure A1, the essays for the three free-response questions are rated on a 9-point scale, and the free-response questions are weighted by 3.0556. This yields the Weighted Section II score for the free-response portion of the AP ELC Exam. The composite score, based on the sum of the weighted scores from sections I and II, is then converted to an AP grade based on the AP Grade Conversion Chart that is shown in Figure A1.

## Illustration

In order to illustrate the potential effects that variation in faculty consultant severity may have on the AP composite scores and AP grades, a small simulation was carried out. Our goal in conducting the simulation was to show, from the vantage point of the student, the "best-case" and "worst-case" scenarios that could occur, given that faculty consultants differ in the level of severity they exercise. The "best-case" scenario for a student would be, by luck of the rater draw, to have three lenient faculty consultants rate the students' three essays. By contrast, the "worst-case" scenario for that student would be to have three severe faculty consultants rate the students' three essays.

The results from the simulation are presented in Table 21. In this example, a hypothetical student with an achievement level of 0.25 logits (theta = 0.250) is rated by three different groups of faculty consultants who were simulated to vary in average severity. Each of the three faculty consultants within a group should be understood to have rated the student's performance on Essays 1 to 3; for example, a1 rated Essay 1, a2 rated Essay 2, …, a8 rated Essay 2, and a9 rated Essay 3. The first group of faculty consultants was lenient (a1 = -0.50, a2 = -0.50, a3 = -0.50). The second group was average—that is, neither lenient nor severe (a4 = 0.00, a5 = 0.00, a6 = 0.00), and third group was severe (a7 = 0.50, a8 = 0.50, a9 = 0.50). (As was shown in the variable map [Figure 2], the vast majority of faculty consultants had severity measures between –0.50 and 0.50 logits. Therefore, when simulating the "lenient" faculty consultant group, we used a severity measure of –0.50 logits. When simulating the "severe" faculty consultant group, we used a severity measure of 0.50 logits.)

Because each group of faculty consultants rated the same student having the same level of English achievement, the scores on the multiple-choice portion of the exam are all the same (Weighted Section I = 84). However, the ratings on the essays for the three free-response questions were designed to vary based on the severity of the group of faculty consultants. For example, a1, the first faculty consultant in the lenient group who scored Essay 1 was expected, based on the many-faceted Rasch model, to assign a rating of 4.301, while a7, the faculty consultant in the severe group who scored the same Essay 1, was simulated to assign a rating of 2.732. The three expected ratings were summed and weighted following the scoring design given in Figure A1.

It is clear that the Weighted Section II scores obtained for the same student from three different faculty consultant groups vary substantially. The AP composite scores range from 81.68 to 124.67, while the AP grades that correspond to these composite scores would be 3, 4, or 5. As will be shown later, the design of the assessment system with faculty consultants nested within essays makes it unlikely that a student would be unlucky enough to receive independent ratings on the three essays from three different faculty consultants who all tend to rate severely. In many instances, differences in faculty consultant severity tend to cancel out, since the student is rated on three essays, not just one. But that may not always be the case.

In the next section of this report, essay ratings and AP grades that have been adjusted for differences in faculty consultant severity are generated for all of the students and compared to the actual ratings and AP grades that students received.

TABLE 21

**Illustration of the Potential Effects of Differences in Faculty Consultant Severity on a Hypothetical Student (Theta = 0.25)**

| | *Lenient (a1=-0.50, a2=-0.50, a3=-0.50)* | *Neither Lenient nor Severe (a4=0.00, a5=0.00, a6=0.00)* | *Severe (a7=0.50, a8=0.50, a9=0.50)* |
|---|---|---|---|
| **WEIGHTED SECTION I** | 84 | 84 | 84 |
| **Free-Response Questions** | | | |
| Essay 1 (Q. 56) | 4.301 | 3.600 | 2.732 |
| Essay 2 (Q. 57) | 4.328 | 3.630 | 2.770 |
| Essay 3 (Q. 58) | 4.681 | 4.002 | 3.247 |
| Sum 13.310 | 11.232 | 8.749 | |
| **(Section II weight)** | (3.0556) | (3.0556) | (3.0556) |
| **WEIGHTED SECTION II** | 40.67 | 34.32 | 26.73 |
| **AP Composite Score** | 124.67 | 118.32 | 81.68 |
| **AP Grade** | 5 | 4 | 3 |

Note.The three faculty consultants within a group should be understood to have rated the student's performance on Essays 1 to 3: for example, a1 rated Essay 1; a2 rated Essay 2; …, a9 rated Essay 3.

TABLE 22

**Impact on Essay Ratings of Adjusting the Ratings for Differences in Faculty Consultant Severity (*N* = 8,642)**

| *Absolute values of the differences between the adjusted and unadjusted ratings* | *Essay 1 (%)* | *Essay 2 (%)* | *Essay 3 (%)* |
|---|---|---|---|
| < .50 | 31.0 | 31.0 | 28.5 |
| >= .50 and < 1.50 | 45.2 | 45.2 | 45.1 |
| >= 1.50 and < 2.50 | 19.1 | 19.2 | 20.8 |
| >= 2.50 and < 3.50 | 4.2 | 4.1 | 4.9 |
| >= 3.50 and < 4.50 | 0.5 | 0.5 | 0.6 |
| >= 4.50 | 0.0 | 0.0 | 0.1 |
| *Summary Statistics (Total)* | | | |
| Mean | 1.02 | 1.02 | 1.06 |
| *SD* | 0.76 | 0.76 | 0.78 |
| *N* | 8,642 | 8,642 | 8,642 |
| Median (Q2) | 0.86 | 0.86 | 0.90 |
| Q1-Q3 (middle 50%) | 0.41-1.47 | 0.41-1.47 | 0.44-1.54 |

## Results for the Total Sample

For each student, we computed an estimate of student achievement adjusted for differences in faculty consultant severity (i.e., an adjusted theta value). We used these theta values to obtain adjusted essay ratings and AP grades. We compared the adjusted essay ratings to operational ratings assigned by faculty consultants to identify instances in which the students' ratings on individual essays would differ if the level of severity of the faculty consultant scoring the essay were taken into account.

Table 22 shows the impact on the essay ratings of adjusting those ratings for differences in faculty consultant severity for the total sample of 8,642 students. (The results reported in this table and in Tables 25 and 26 are based on the absolute values of the differences between the adjusted and unadjusted ratings.) For Essays 1 and 2, in approximately 76 percent of the cases in which the adjusted and unadjusted ratings differed, those differences were less than 1.50 points on the 9-point scale. Similarly, for Essay 3, in approximately 74 percent of the cases in which the adjusted and unadjusted ratings differed, those differences were less than 1.50 points. The mean differences between adjusted and unadjusted ratings were 1.02, 1.02, and 1.06 points respectively for Essays 1 to 3. These data indicate that the average difference between the unadjusted and adjusted ratings across essays was approximately one score point.

Table 23 shows the impact on AP grades of adjusting the essay ratings for differences in faculty consultant severity. (The table is based on the absolute values of the differences between the adjusted and unadjusted AP grades.) Approximately 70 percent of the students would have received the same AP grade whether or not their essay ratings were adjusted for the level of severity that the faculty consultants scoring their essays exercised. Almost all (99.7 percent) of the differences were one grade or less. The average difference between the adjusted and unadjusted AP grades is 0.30 (*SD* = 0.46). Table 24 presents a cross-tabulation of student AP grades unadjusted for faculty consultant severity and AP grades adjusted for differences in faculty consultant severity.

## Results by Student Subgroup

In addition to examining the impact on AP grades of adjusting essay ratings for differences in faculty consultant severity for the total sample, it is important to consider the impact of those adjustments on the AP grades of different subgroups of students. The specific subgroups examined here are based on student gender, race/ethnicity, and a self-report response regarding

TABLE 23

**Impact on AP Grades of Adjusting Essay Ratings for Differences in Faculty Consultant Severity (*N* = 8,642)**

| Changes in AP Grade | AP Grade % |
|---|---|
| 0 (same grade) | 70.1 |
| 1 | 29.6 |
| 2 | 0.3 |
| *Summary Statistics (Total)* | |
| Mean | .30 |
| *SD* | .46 |
| *N* | 8,642 |
| Median (Q2) | 0 |
| Q1-Q3 (middle 50%) | 0-1 |

Note. This table is based on the absolute values of the differences between the adjusted and unadjusted AP grades.

TABLE 24

**Cross-Tabulation of AP Grades Adjusted and Unadjusted for Differences in Faculty Consultant Severity**

| *AP Grade Unadjusted for Differences in Faculty Consultant Severity* | *AP Grade Adjusted for Differences in Faculty Consultant Severity* | | | | | |
|---|---|---|---|---|---|---|
| | | *1* | *2* | *3* | *4* | *5 Total* |
| 1 | 344 (91.2%) | 33 | 0 | 0 | 0 | 377 |
| 2 | 204 | 1727 (76.3%) | 329 | 2 | 0 | 2262 |
| 3 | 0 | 500 | 2004 (66.2%) | 513 | 11 | 3028 |
| 4 | 0 | 3 | 441 | 1165 (60.3%) | 324 | 1933 |
| 5 | 0 | 0 | 9 | 214 | 819 (78.6%) | 1042 |
| Total | 548 | 2263 | 2783 | 1894 | 1154 | 8642 |

what language the student knows best. Table 25 shows the impact of the adjustments on the AP grades of these subgroups. Although there is some variability across subgroups, there does not appear to be any strong evidence that the changes in the essay ratings would impact some subgroups more than others. The impact would be the strongest on the subgroup that reports a language other than English as the one that they know best (EBL = 3). The mean differences between adjusted and unadjusted ratings for this subgroup are 1.42, 1.41, and 1.21 for the three essays. Caution should be exercised in generalizing from these results, however, because the sample size is quite small ($N$ = 36).

TABLE 25

**Impact on Essay Ratings of Adjusting for Differences in Faculty Consultant Severity by Student Subgroup**

| | *Sample* | | *Mean Differences between Adjusted and Unadjusted Ratings* | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | *Essay 1* | | *Essay 2* | | *Essay 3* | |
| | *Freq.* | *Percent* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| ***Gender Subgroups*** | | | | | | | | |
| 1 Female | 5,574 | 64.5 | 1.01 | .76 | 1.01 | .75 | 1.05 | .77 |
| 2 Male | 3,068 | 35.5 | 1.02 | .78 | 1.02 | .78 | 1.08 | .79 |
| ***Race/Ethnicity Subgroups*** | | | | | | | | |
| 1 American Indian or Alaska Native | 42 | 0.5 | 1.05 | .91 | 1.04 | .90 | .98 | .74 |
| 2 Black or African American | 410 | 5.1 | 1.01 | .74 | 1.01 | .74 | .99 | .77 |
| 3 Mexican American or Chicano | 266 | 3.3 | .98 | .78 | .97 | .77 | 1.00 | .76 |
| 4 Asian, Asian American, or Pacific Islander | 840 | 10.4 | 1.02 | .74 | 1.02 | .74 | 1.04 | .73 |
| 5 Puerto Rican | 55 | 0.7 | 1.05 | .77 | 1.05 | .77 | 1.04 | .82 |
| 6 South American, Latin American, Central American, or other Hispanic | 194 | 2.4 | .99 | .76 | .99 | .76 | .99 | .77 |
| 7 White | 5,997 | 74.3 | 1.01 | .77 | 1.01 | .76 | 1.07 | .78 |
| 8 Other | 266 | 3.3 | .98 | .77 | .99 | .77 | 1.01 | .81 |
| ***Best Language Subgroups*** | | | | | | | | |
| 1 English | 7,894 | 93.7 | 1.02 | .76 | 1.02 | .76 | 1.07 | .78 |
| 2 English and another language about the same | 491 | 5.8 | 1.01 | .74 | 1.01 | .74 | 1.01 | .71 |
| 3 Another language | 36 | 0.4 | 1.42 | .95 | 1.41 | .94 | 1.21 | .88 |
| TOTAL | 8,642 | 100.0 | 1.02 | .76 | 1.02 | .76 | 1.06 | .78 |

Note. The values in this table are based on the absolute values of the differences between the adjusted and unadjusted AP grades.

TABLE 26

**Impact on AP Grades of Adjusting Essay Ratings for Differences in Faculty Consultant Severity (Student Subgroups)**

| | *Sample* | | *Changes in AP Grades* | |
|---|---|---|---|---|
| | *Freq.* | *Percent* | *Mean* | *SD* |
| ***Gender Subgroups*** | | | | |
| 1 Female | 5,574 | 64.5 | .30 | .46 |
| 2 Male | 3,068 | 35.5 | .30 | .47 |
| ***Race/Ethnicity Subgroups*** | | | | |
| 1 American Indian or Alaska Native | 42 | 0.5 | .17 | .38 |
| 2 Black or African American | 410 | 5.1 | .25 | .44 |
| 3 Mexican American or Chicano | 266 | 3.3 | .26 | .44 |
| 4 Asian, Asian American, or Pacific Islander | 840 | 10.4 | .32 | .48 |
| 5 Puerto Rican | 55 | 0.7 | .27 | .45 |
| 6 South American, Latin American, Central American, or other Hispanic | 194 | 2.4 | .26 | .44 |
| 7 White | 5,997 | 74.3 | .31 | .47 |
| 8 Other | 266 | 3.3 | .30 | .46 |
| ***Best Language Subgroups*** | | | | |
| 1 English | 7,894 | 93.7 | .30 | .47 |
| 2 English and another language about the same | 491 | 5.8 | .25 | .44 |
| 3 Another language | 36 | 0.4 | .39 | .55 |
| TOTAL | 8,642 | 100.0 | .30 | .46 |

Note. The values in this table are based on the absolute values of the differences between the adjusted and unadjusted AP grades.

Turning now to the impact of adjustments on AP grades (see Table 26), we see again that adjusting essay ratings for differences in faculty consultant severity does not appear to differentially impact student subgroups. For most of the subgroups, the mean differences between the adjusted and unadjusted AP grades are around the average of 0.30. The only subgroup with a somewhat larger mean difference (0.39) was the small group ($N$ = 36) of students who reported a language other than English was their best language.

## *Summary of Results in Terms of the Research Questions*

In this section of the report, we briefly summarize results from our analyses that pertain to each of the research questions framing this investigation.

1. Do faculty consultants differ in the levels of severity they exercise when scoring students' essays written for Section II of the 1999 AP English Literature and Composition Exam? What is the best approach for calibrating faculty consultants?

The faculty consultants differed somewhat in the levels of severity they exercised. The results of the chi-square analysis indicated that the overall difference in severity between faculty consultants was statistically significant. The separation index was 1.57, indicating that within the sample of 605 faculty consultants, there were about one-and-a-half statistically distinct strata of severity. The faculty consultants were somewhat more lenient when rating essays for Question 58 than when rating essays for Questions 56 and 57.

We considered three approaches to calibrating faculty consultants. We decided against using the first two approaches given the sheer volume, computational effort, and costs associated with implementing them. The approach we adopted involved calibrating the multiple-choice questions, free-response questions, and faculty consultants simultaneously so that the faculty consultants were linked through the multiple-choice questions.

2. Are there interactions between faculty consultant severity and extraneous student background characteristics (e.g., gender, race/ethnicity, and best language) that may impact essay ratings and grades on the 1999 AP English Literature and Composition Exam?

The results from our analyses indicate that there were not statistically significant group-level differential severity/leniency effects related to student gender, race/ethnicity, or best language. However, there were individual faculty consultants who appeared to have displayed differential severity/leniency effects related to each background characteristic. Our analyses identified 18 faculty consultants who exhibited differential

severity/leniency effects related to gender, 118 who exhibited differential severity/leniency effects related to race/ethnicity, and 3 faculty consultants who exhibited differential severity/leniency effects related to best language. It is important to emphasize that, in the case of student race/ethnicity and best language, each of the faculty consultants the analysis identified as "biased" in their ratings rated only a small number of essays composed by students from the subgroups in question. Consequently, the designation of these faculty consultants as "biased" is very preliminary and would need further verification by having the faculty consultant rate more essays written by students from these particular subgroups to obtain a larger, more representative sample of his or her rating behavior upon which to base judgments regarding possible bias.

3. Do adjustments for faculty consultant severity have an impact on essay ratings and/or on AP grades?

If students' essay ratings were adjusted for the level of severity the faculty consultants exercised in scoring those essays, then for Essays 1 and 2, in approximately 76 percent of the cases in which the adjusted and unadjusted ratings differed, those differences would have been less than 1.50 points on the 9-point scale. Similarly, for Essay 3, in approximately 74 percent of the cases in which the adjusted and unadjusted ratings differed, those differences would have been less than 1.50 points. The mean differences between adjusted and unadjusted ratings were 1.02, 1.02, and 1.06 points respectively for Essays 1 to 3. The average difference between the unadjusted and adjusted ratings across essays was approximately one score point.

If students' AP grades were adjusted for the level of severity the faculty consultants exercised in scoring their essays, then approximately 70 percent of the students would have received the same AP grade whether or not their essay ratings were adjusted for the level of severity that the faculty consultants scoring their essays exercised. Almost all (99.7 percent) of the differences were one grade or less. The average difference between the adjusted and unadjusted AP grades was 0.30 ($SD$ = 0.46).

4. Does faculty consultant severity differentially impact essay ratings and/or AP grades for student subgroups based on student gender, race/ethnicity, or best language?

Although there was some variability across subgroups, there did not appear to be any strong evidence that adjusting essay ratings for faculty consultant severity would impact some subgroups more than others. The impact would be the strongest on the subgroup that reported a language other than English as the one that they knew best (EBL = 3). The mean differences between adjusted and unadjusted ratings for this subgroup were 1.42, 1.41, and 1.21 for the three essays. Caution should be exercised in generalizing from these results, however, because the EBL = 3 subgroup was composed of only 36 students.

If students' AP grades were adjusted for the level of severity the faculty consultants exercised in scoring their essays, there does not appear to be strong evidence that such adjustments would impact some subgroups more than others. For most of the subgroups, the mean differences between the adjusted and unadjusted AP grades were around the average of 0.30. The small subgroup of students ($N$ = 36) whose best language was a language other than English (EBL = 3) showed the largest mean difference (0.39).

# Discussion

The findings from the present study confirm those of Coffman and Kurfman (1968), Braun (1988), Longford (1994a, 1994b), Myford and Mislevy (1995), and Bridgeman, Morgan, and Wang (1996). In all six studies, faculty consultants scoring AP Examinations differed in the levels of severity they exercised; however, it is important to note that differences between faculty consultants were more pronounced for some exams than for others. Students taking the AP ELC Exam write essays for three free-response questions. The results from our study and from earlier studies (Braun, 1988; Bridgeman, Morgan and Wang, 1996) indicate that faculty consultants scoring the essays appearing on the AP ELC Exam are not all interchangeable. Results from the present study suggest that there were about one-and-a-half statistically distinct strata of severity within the 605 faculty consultants included in our analyses.

To produce a raw composite score on the AP ELC Exam, each student's ratings on the three essays are combined and weighted. Since three faculty consultants contribute to the composite score, some (but not all) of the differences in faculty consultant severity tend to cancel out when the ratings are combined. While the differences in faculty consultant severity are not large, adjusting for these differences could impact the AP grades of a sizable number of the students taking the AP ELC Exam. The results from our study indicate that if students' ratings had been adjusted for severity differences that remained after the three ratings were combined and weighted, the AP grades of about 30 percent of the students would have been different from the one they received. In about half of these cases, the student's

AP grade would have been one grade higher. In the other half of the cases, the student's AP grade would have been one grade lower. Additionally, there would have been some students whose AP grades would have been two grades higher (or lower).

The issue of whether to adjust ratings for faculty consultant severity differences in the scoring of the essays is particularly critical for students whose AP grades lie near critical cut points in the score distribution. Adjusting ratings for faculty consultant severity differences would negatively impact some of these students and positively impact others. For example, suppose a college has a policy of granting college credit to students who receive an AP grade of 3 or higher on the AP ELC Exam. Some students who took the exam may have an AP grade of 2 that with adjustment for faculty consultant severity differences would result in their receiving an AP grade of 3 rather than 2. Adjusting for faculty consultant severity differences would make these students eligible to receive credit at this college; without adjustment, they would be ineligible. Alternatively, some students may have an AP grade of 3 that with adjustment for faculty consultant severity differences would result in their receiving an AP grade of 2 rather than 3. Without adjustment for faculty consultant severity, they would be eligible for college credit; with adjustment, they would become ineligible.

As Braun and Wainer (1989) pointed out, the reactions of students, parents, teachers, and school administrators to statistical adjustment of ratings need to be taken into consideration when policymakers grapple with the issue of whether or not to implement a statistical adjustment procedure. In our experience, often these audiences are quick to acknowledge that when students' essays were scored by severe faculty consultants, adjusting their ratings upward leads to fairer AP grades, since students should not be penalized by the "luck of the rater draw." However, these same audiences frequently balk at the notion of adjusting ratings downward, failing to acknowledge that such adjustments are equally necessary in those cases in which the faculty consultants who scored the students' essays were more lenient than other faculty consultants.

If adjustments in ratings were to be made for faculty consultant differences, those adjustments would need to be based on trustworthy, stable measures of faculty consultant performance. Several important questions arise: Just how variable is individual faculty consultant performance from day to day, from morning to afternoon, from essay folder to essay folder? Is it appropriate to adjust ratings based on a single calibrated measure of each faculty consultant's severity (i.e., calculate a faculty consultant severity measure based on the faculty consultant's ratings of all essays he/she scored during an AP reading and then use those measures as the basis for adjusting students' ratings)? As an alternative, should the adjustment procedure take into consideration not only the level of severity of the particular faculty consultants who scored the essay but also the day the essay was scored? As Braun and Wainer (1989) suggested, the more appropriate strategy might be to remove the chance variation from students' ratings that is due to systematic differences between faculty consultants and between days (i.e., calibrate both faculty consultants and days to eliminate these two sources of unwanted variation in the ratings). Longford (1994a) proposed yet a third alternative: Adjust student ratings for differences not only in faculty consultant severity but also in faculty consultant consistency.

Unfortunately, the results from the present study do not help policymakers decide which of these three statistical adjustment procedures to employ. Nor do the results provide evidence that would help policymakers determine whether it is wise at this stage to adopt a policy of statistically adjusting ratings for systematic differences in faculty consultant performance. In our view, before such decisions can be made, it will be important to look at the stability and consistency of individual faculty consultant performance over time. The Advanced Placement Research and Development Committee has recently approved a proposal to investigate the stability and consistency of individual faculty consultant performance as an AP reading progresses (Wolfe, Engelhard, and Myford, 2001). Specifically, the researchers will determine whether AP ELC faculty consultants exhibit changes over time in their levels of severity, accuracy, and use of individual categories on the scoring guidelines. The follow-up study should provide results that can help inform policy decision making.

## *Implementing Statistical Adjustment Procedures—Feasibility Issues*

Policymakers will also need to consider the changes that would need to be made regarding the way that AP readings are carried out in order to provide the data necessary to implement the various statistical adjustment procedures that have been proposed. To collect the data needed to implement a particular procedure, certain requirements must be met. These requirements differ from procedure to procedure, as we shall see.

To carry out the adjustment procedure that Braun (1988) advocated, a carefully designed calibration experiment is necessary that involves selecting a small

sample of essays and faculty consultants and controlling the order in which those essays were read. (Braun used 32 essays and 12 AP ELC faculty consultants. In his study, the allocation plan of essays to faculty consultants involved having each faculty consultant rate eight essays each day, and having each essay read three times on each of four consecutive days.) The implementation of a partially balanced incomplete block design is a delicate balancing act that requires each pair of faculty consultants involved in the experiment to read either two essays in common or no essays in common (see Braun [1988], page 5, figure 1, for the layout of his allocation plan). The information obtained from the small experiment becomes the basis for adjusting ratings for the entire pool of essays, even though only a small subset of faculty consultants and essays are actually involved in the experiment. The careful allocation of faculty consultants to essays is necessary in order to estimate certain variance components. (It should be noted that given recent developments in generalizability theory [Brennan, 2001], the requirements for data designs have been relaxed somewhat, which today may makes it less burdensome to collect the data necessary to compute the variance components using this least-squares regression approach.)

One question that may arise concerning the operational implementation of Braun's approach is the extent to which the estimates of faculty consultant severity that one would obtain are accurate measures, especially for the faculty consultants who were not included in the small experiment. When Braun conducted his experiment in 1985, the total number of AP ELC faculty consultants was 273, and the total number of students taking the AP ELC Exam was 75,705. By comparison, in 1999 there were 612 AP ELC faculty consultants, and the total number of students taking the AP ELC Exam was close to 175,000. As Braun and Wainer (1989) noted, before deciding to implement their procedure operationally, it would be important to compare the calibration obtained from analyzing the full dataset (i.e., all the ratings from the operational reading) to the calibration obtained from the small experiment to determine the extent to which the results were similar. If Braun's experiment were repeated today using just 12 AP ELC faculty consultants, how dependable would the measures of faculty consultant severity be for the remaining 600 faculty consultants? How accurate would the adjustments in the ratings be for the 175,000 students' essays not included in the experiment?

To carry out the statistical adjustment procedure that Longford (1994a) advocated, it is necessary to conduct several faculty consultant reliability studies as part of each AP reading. That is, a random sample of essays from the reading would each need to be scored by two faculty consultants. Furthermore, a faculty consultant reliability study would need to be carried out for each and every free-response question included on the AP Exam. (Current policy is to have a single faculty consultant score each essay in the operational reading, and faculty consultant reliability studies are only conducted on a periodic basis, not as a part of each operational reading.) The double scoring is necessary in order to compute the faculty consultant inconsistency variance component. The inconsistency variance component from the reliability study is then used to impute an estimate of the inconsistency variance component for the operational ratings of essays for that free-response question. Longford (1994a) argued that "once the inconsistency variance can be imputed, based on the estimates from a large number of previous forms of the same test, the reliability studies can be dispensed with" (p. 13). This may prove problematic for the AP Program, since test forms are never repeated. Each year the previous year's exam is fully disclosed, and test forms are not formally equated from year to year. Therefore, some might question whether using faculty consultant invariance components estimated from one or more previous years' exam topics provides a fair and justifiable basis for adjusting students' ratings for differences in faculty consultant severity and inconsistency for ratings given in the present year. If the AP Program did not want to use faculty consultant invariance components estimated from previous years' exams, then it would be necessary to conduct a faculty consultant reliability study for each free-response question included on the present year's exam. That would involve much additional expense for the AP Program to pick up in order to double score a sample of essays for each free-response question.

Another potential problem with the Longford adjustment procedure is that the adjustments in ratings that were made for faculty consultants having small workloads (i.e., those rating few essays) were clearly different from the adjustments in ratings made for faculty consultants having large workloads. There was much more uncertainty associated with the adjustments in ratings made for faculty consultants who rated few essays, Longford noted. To combat this problem, Longford recommended that the reading be structured so that no faculty consultants would have small workloads. Is that is a realistic goal for AP readings, or is likely that there will continue to be faculty consultants who will have small workloads? If faculty consultants are likely to continue to vary in terms of their workloads, then it will be necessary to take that into consideration when deciding whether this adjustment procedure would adequately deal with that problem. It would be important to be

able to show that statistical adjustments made to ratings of students whose essays were judged by faculty consultants who had small workloads were just as trustworthy as the adjustments made to ratings of students whose essays were judged by faculty consultants who had large workloads.

To be sure, the adjustment procedure that we used in this study also has shortfalls. One of the biggest challenges that we face is being able to include all the rating data from an AP ELC reading in a single analysis. The version of the FACETS software that we employed in this study could not handle a data set this large. Consequently, we chose to work with a 5 percent sample in this study. Clearly, before the AP Program could ever consider using FACETS operationally, there would need to be a version of the software that could accommodate the large candidate volumes that are typically encountered in AP readings. The recently approved research proposal previously cited (Wolfe, Engelhard, and Myford, 2001) will provide a test of whether the newest version of FACETS, which claims to be able to analyze much larger data sets, is up to this task.

A second concern we have with using FACETS to statistically adjust students' essay ratings is establishing sufficient connectivity among all the faculty consultants who participate in an AP reading. Allocating faculty consultants to essays must result in a network of links that is complete enough to connect all the faculty consultants through common essays if the faculty consultants are to be directly compared in terms of their severity (Engelhard, 1997; Lunz, Wright, and Linacre, 1990). Otherwise, ambiguity in interpretation results. When there are disconnected subsets of faculty consultants in a FACETS analysis, only faculty consultants that appear in the same subset can be directly compared. Attempts to compare faculty consultants that appear in two or more different disconnected subsets can be misleading. In this study, we calibrated the multiple-choice questions, free-response questions, and faculty consultants simultaneously, linking the faculty consultants through the multiple-choice questions.

While the current study has demonstrated that it is possible to link all the faculty consultants through the multiple-choice questions, we are planning to experiment with a more direct method of establishing connectivity in our follow-up study (Wolfe, Engelhard, and Myford, 2001). In this study, we will introduce "benchmark essays" into the reading (i.e., a set of essays that a group of expert readers have previously rated in order to obtain a consensus rating for each essay). We say "more direct" because connecting faculty consultants through students' responses to multiple-choice questions only allows the AP Program to obtain indirect measures of each faculty consultant's performance (i.e., an indication of whether the ratings the faculty consultant gives students' essays are "in sync" with the overall level of performance each student displayed on the multiple-choice section of the exam). Linking faculty consultants through their ratings of common benchmark essays would provide the AP Program with another option that does not rely on the multiple-choice questions as the only means available to establish faculty consultant connectivity. Additionally, introducing benchmark essays into a reading would allow the AP Program to monitor each faculty consultant's level of accuracy over time, comparing a faculty consultant's performance to known standards of performance (i.e., the expert readers' consensus ratings of the benchmark essays).

A third concern we would raise based on results from the present study is whether it is appropriate to adjust ratings for differences in faculty consultant severity if those adjustments are based on a single measure of severity for each faculty consultant. As Myford and Wolfe argue (2001), a faculty consultant severity effect can present itself in several ways, some more subtle than others. Some severe faculty consultants may underestimate the level of student performance across the entire achievement continuum. These faculty consultants do not accurately assess the level of achievement of students at any point along the continuum. Rather, they tend to consistently assign ratings that are lower than the ratings that other faculty consultants would assign the same students. When researchers use the term "severity effect," it is often with this intended meaning. However, there are other more subtle ways in which faculty consultants may exhibit a severity effect. For example, some faculty consultants may exhibit a tendency to cluster their ratings around a particular category on a rating scale (i.e., show restriction of range in their ratings). That category may be at the high end of the scale, the low end of the scale, or in the middle of the scale. If a faculty consultant's ratings tend to cluster at the lower end of the scale, then that may signal severity. Note that in this example, the faculty consultant does not underestimate student achievement across the entire achievement continuum—only along a portion of that continuum. The net effect is still detectable as faculty consultant severity, though the pattern of ratings for a faculty consultant showing restriction of range may differ somewhat from the pattern of ratings for a faculty consultant who consistently assigns lower ratings than other faculty consultants to all students. However, as these examples illustrate, it is often difficult to differentiate clearly between restriction of range and severity as separate faculty consultant effects.

Finally, a faculty consultant may selectively exhibit a severity effect. That is, a faculty consultant may be differentially severe, showing a tendency to assign ratings that are lower than expected to essays composed by certain subgroups of students, given the ratings that other faculty consultants assign these students' essays. However, the faculty consultant may not show this same tendency when rating essays composed by other subgroups of students. This is a more subtle form of the faculty consultant severity effect that might be referred to as "differential severity." The key question that arises, then, is this: Is it appropriate to use a "one-size-fits-all" approach to adjusting ratings for faculty consultant severity differences if, indeed, faculty consultant severity differences can present in these categorically different ways? Perhaps what is needed is a more mathematically sophisticated approach to adjustment that would take into account the potentially localized nature of a faculty consultant severity effect. Such an approach would not make the assumption that a severe faculty consultant exercises a constant level of severity, no matter what student he or she is rating, no matter what day the rating occurs, no matter whether the essay is scored in the morning or in the afternoon, no matter whether the essay is the first in the folder or the last, etc. Rather, this alternative approach would take into consideration these contextualized (and potentially powerful) facets of the reading and use information about differences in faculty consultant performance related to these facets in adjusting ratings. Accordingly, additional research is needed to determine the best way to interpret interaction effects and their impact on adjustment for faculty consultant severity.

## *Quality Control Monitoring Using a Many-Faceted Rasch Measurement Approach*

The goal of the AP Program is to develop an examination system that provides a framework for drawing valid, reliable, and fair inferences regarding the level of achievement that students taking AP Exams have attained. AP grades should lead to valid generalizations about students' levels of achievement in the content domains (Linn, Baker, and Dunbar, 1991). Ideally, a student's AP grade should not be tied to the particular faculty consultants who evaluated the student's essays, to the particular set of multiple-choice and free-response questions included on the exam, to the scoring guidelines that were employed to judge the essays, or to the particular students who took the AP Exam. A student's AP grade should be invariant across the specific details of the AP Examination system.

If the goal is to make valid inferences from a student's AP grade, then psychometric models are needed that help determine how well the various aspects of the examination system are operating. Earlier in this paper, we described a conceptual model for the measurement of English achievement on the AP ELC Exam. The conceptual model contained a number of intervening variables that, if not monitored closely, have the potential to threaten the validity of the AP grades by introducing unwanted sources of construct-irrelevant variance into the examination process. The psychometric model we employed in this study incorporated some (but not all) of the possible intervening variables. We used FACETS, an item response theory (IRT) model-based measurement approach, to analyze data from the 1999 AP ELC Exam.

In our FACETS analyses, we examined how faculty consultants, students, test questions, and scoring guidelines in the AP ELC examination system performed. Using selected pieces of output from our analyses, we were able to pinpoint aspects of the examination system that were functioning as intended, as well as potentially problematic aspects. The analyses provided specific information about how each "element" of each facet (i.e., each faculty consultant, student, and test question) within the examination system was performing—detailed information that those in charge of monitoring quality control for the AP ELC Examination could use to initiate meaningful changes to improve the system.

In general, results from our study indicate that the AP ELC examination system is functioning quite well. The 58 questions included on the 1999 AP ELC Exam work together to define a unidimensional English achievement construct. There seems to be little evidence of multidimensionality in the data, and only one of the questions (Question 20) appears to function in a somewhat redundant manner. (This question was one of the easiest on the exam, which suggests that the problem with this question may be one of ineffective examinee targeting rather than dependency [J.M. Linacre, personal communication, March 14, 2002.]) Overall, the scores on the multiple-choice questions correspond well to the ratings on the free-response questions. Therefore, it appears that scores on the multiple-choice questions can be meaningfully combined with the ratings on the free-response questions to produce a single summary measure (i.e., the AP grade) that can appropriately capture the essence of student performance across the 58 questions. The set of 58 questions succeeds in defining about two-and-a-half statistically distinct levels of English achievement among the students. Overall, most student score profiles show consistent performance across the 58 questions. While there is some variation in the level of severity that individual faculty consultants exercise, most used the 9-category rating scale in a consistent fashion. Additionally, most faculty consultants maintained a uniform level of severity when rating

essays of subgroups of students having different background characteristics.

What changes could those in charge of monitoring quality control for the AP ELC Exam initiate to fine-tune their system? Results from the study point to several specific steps that could be taken to improve the system, even if the AP Program had no intention of using FACETS results to adjust composite scores or AP grades for differences in faculty consultant severity. About 4 percent of the students had score profiles that included one or more highly unexpected or surprising responses to multiple-choice questions and/or unexpected ratings on the essays composed for the free-response questions, given the faculty consultant's level of severity, the student's performance on the multiple-choice questions, and the other ratings that the student received on his/her essays. Before issuing score reports, those monitoring quality control for the AP ELC Examination could use the student fit statistics that FACETS provides to identify misfitting students. They could then examine each misfitting student's score profile (using quality control tables or charts, like the examples we provided in this report) to determine whether the student's AP grade provides a valid indicator of the student's overall level of English achievement and should be left to stand as it is, or perhaps whether one or more scores on individual questions are aberrant and require further investigation prior to sending out the score report.

Similarly, FACETS identified a small number of faculty consultants as misfitting. By reviewing the detailed information provided in quality control charts and/or tables for misfitting faculty consultants, the table leaders could gain an understanding of the specific nature of each faculty consultant's misfit. The table leaders would then be in a much stronger position to determine how best to work with each faculty consultant, providing individually targeted retraining to help them learn to use the scoring guidelines in a more consistent fashion. Further, if FACETS analyses could be conducted in "real time" (i.e., while an AP reading is taking place), then table leaders could use the faculty consultant fit statistics to identify early on those faculty consultants who need additional training before they are allowed to score operationally.

Table leaders could also make use of the results from the FACETS bias analyses to help them devise targeted retraining activities. The output from these analyses could help them identify individual faculty consultants who showed a differential severity effect related to a specific student background characteristic (e.g., student gender, race/ethnicity, or best language). By reviewing bias analyses quality control charts and/or tables for each faculty consultant, table leaders could pinpoint the particular students most affected. Additionally, the table leaders could identify specific patterns of differential severity and then use that information to determine whether there are groups of faculty consultants who exhibit similar patterns. If so, then retraining activities could be tailored to meet their specific needs, helping faculty consultants to become aware of the biases they exhibit and exploring together positive steps they could take to deal with those biases.

# References


Advanced Placement Program (2001). *Technical manual*. Retrieved May 23, 2001 from the World Wide Web: http://www.collegeboard.com/ap/techman.

American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 357–374.

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

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park: Sage.

Andrich, D. (1998). Thresholds, steps, and rating scale conceptualization. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 12(3), 648.

Beale, E.M.L. & Little, R.J.A. (1975). Missing data in multivariate analysis. *Journal of the Royal Statistical Society* (B), 129–145.

Bernardin, H.J. & Pence, E.C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology*, *65*, 60–66.

Bleistein, C. & Maneckshana, B. (1995). *English literature and composition folder study* (Unpublished Statistical Report No. SR-95-39). Princeton, NJ: Educational Testing Service.

Braun, H.I. (1986). *Calibration of essay readers. Final report* (ETS Program Statistics Research Technical Report No. 86-68). Princeton, NJ: Educational Testing Service.

Braun, H.I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, 13(1), 1–18.

Braun, H.I. & Wainer, H. (1989). *Making essay test scores fairer with statistics* (ETS Program Statistics Research Technical Report No. 89–90). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED395028)

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

Bridgeman, B. Morgan, R. & Wang, M. (1996). *Reliability of Advanced Placement Examinations* (ETS Research Report RR-96-3). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED400331)

Campbell, E.H. (1993, March). *Fifteen raters rating: An analysis of selected conversation during a placement rating session.* Paper presented at the 44th Annual Meeting of the Conference on College Composition and Communication, San Diego, CA. (ERIC Document Reproduction Service No. ED358465)

Cason, G.J. & Cason, C.L. (1984). A deterministic theory of clinical performance rating. *Evaluation and the Health Professionals, 7*, 221–247.

Chase, C. I. (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement,* 23(1), 33–41.

Coffman, W.E. & Kurfman, D. (1968). A comparison of two methods of reading essay examinations. *American Educational Research Journal, 5*(1), 99–107.

The College Board (1999a). *Released exam—1999 AP English Literature and Composition*. New York: Author.

The College Board (1999b). 1999 *Advanced Placement Test Analyses*, *Forms 3VBP*. Unpublished statistical report. Princeton, NJ: Educational Testing Service.

The College Board (2000). *Advanced Placement Program Course Description—English Language and Composition, English Literature and Composition.* New York: Author.

The College Board and Educational Testing Service (1994). *College and university guide to the Advanced Placement Program.* New York: Authors.

Crocker, L. & Algina, J. (1986). *Introduction to classical & modern test theory.* Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

DeGruijter, D.N.M. (1984). Two simple models for rater effects. *Applied Psychological Measurement,* 8(2), 213–218.

Du, Y. & Wright, B.D. (1997). Effects of student characteristics in a large-scale direct writing assessment. In M. Wilson, G. Engelhard, Jr. & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 1–24). Stamford, CT: Ablex Publishing Company.

Ebel, R.L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407–424.

Engelhard, G. Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*(3), 171–191.

Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement,* 31(2), 93–112.

Engelhard, G. Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement,* 33(1), 56–70.

Engelhard, G. Jr. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement,* 1(1), 19–33.

Engelhard, G. Jr. (in press). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis.* Mahway, NJ: Lawrence Erlbaum Associates, Pub.

Engelhard, G. Jr., Gordon, B. & Gabrielson, S. (1992). The influences of mode of discourse, experimental demand, and gender on the quality of student writing. *Research in the Teaching of English,* 26(3), 120–142.

Engelhard, G., Jr., Gordon, B., Siddle-Walker, E.V. & Gabrielson, S. (1994). Writing tasks and gender: Influences on writing quality of black and white students. *The Journal of Educational Research,* 87(4), 197–209.

Engelhard, G., Jr., Myford, C.M. & Cline, F. (2000). *Investigating assessor effects in NBPTS assessments for Early Childhood/Generalist and Middle Childhood/Generalist* (ETS Research Report RR-00-13). Princeton, NJ: Educational Testing Service.

Fisher, W.P. Jr. (1992). Reliability statistics. *Rasch Measurement: Transactions of the Rasch Measurement SIG,* 6(3), 238.

Fitzpatrick, A.R., Ercikan, K., Yen, W.M. & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education,* 11(2), 195–208.

Fleiss, J.L. (1981). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement, 5*, 105–112.

Garner, M., & Engelhard, G., Jr. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education,* 12(10), 29–51.

Gordon, B. & Engelhard, G., Jr. (1995, October). *How accurately can readers identify the gender of student writers?* Paper presented at the annual meeting of the Georgia Educational Research Association, Atlanta, GA.

Graham S., & Dwyer, A. (1987). *Effects of the learning disability label, quality of writing performance, and examiner's level of expertise on the evaluation of written products.* (ERIC Document Reproduction Service No. ED294351)

Gyagenda, I.S., & Engelhard, G. Jr. (1998, April). *Rater, domain, and gender influences on the assessed quality of student writing using weighted and unweighted scoring.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA. (ERIC Document Reproduction Service No. ED422349)

Houston, W.M., Raymond, M.R. & Svec, J.C. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement,* 15(4), 409–421.

Lance, C.E., LaPointe, J.A. & Stewart, A.M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology,* 79(3), 332–340.

Linacre, J.M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J.M. (1998). *A user's guide to FACETS: Rasch measurement computer program.* Chicago: MESA Press.

Linacre, J.M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement,* 3(2), 103–122.

Linacre, J.M., Engelhard, G., Jr., Tatum, D.S. & Myford, C.M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research,* 21(6), 569–577.

Linn, R.L., Baker, E. & Dunbar, S.B. (1991). Complex performance-based assessments: Expectations and validation criteria. *Educational Researcher,* 20(8), 15–21.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Longford, N.T. (1993). *Reliability of essay rating and score adjustment* (ETS Technical Report No. 93-36). Princeton, NJ: Educational Testing Service.

Longford, N.T. (1994a). *A case for adjusting subjectively rated scores in the Advanced Placement tests.* (ETS Program Statistics Research Technical Report No. 94-5). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED380502)

Longford, N. T. (1994b). Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics*, *19*, 171–201.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lumley, T. & McNamara, T.F. (1993, August). *Rater characteristics and rater bias: Implications for training.* Paper presented at the Language Testing Research Colloquium, Cambridge, England, UK. (ERIC Document Reproduction Service No. ED365091)

Lunz, M.E. & Stahl, J.A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13*, 425–444.

Lunz, M.E. Stahl, J.A., & Wright, B.D. (1996). The invariance of rater severity calibrations. In G. Engelhard, Jr., & M. Wilson (Eds.), *Objective Measurement: Theory into Practice* (Vol. 3, pp. 99–112). Norwood, NJ: Ablex.

Lunz, M.E. Wright, B.D. & Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education,* 3(4), 331–345.

McArthur, D.L. (1981). *Bias in the writing of prose and its appraisal* (CSE Report No. CSE-RR-162). Los Angeles, CA: Center for the Study of Evaluation. (ERIC Document Reproduction Service No. 217073)

McDaniel, B.A. (1985, March). *Ratings vs. equity in the evaluation of writing.* Paper presented at the 36th Annual Meeting of the Conference on College Composition and Communication, Minneapolis, MN. (ERIC Document Reproduction Service No. ED260459)

McNamara, T. (1996). *Measuring second language performance.* New York: Longman.

Maneckshana, B. Morgan, R., & Batleman, M. (1999, November). *Advanced Placement English literature and composition form 3VBP reader reliability study.* Unpublished statistical report, Educational Testing Service, Princeton, NJ.

Morgan, R. (1998). *An examination of the impact of folder position and the reading day on the scoring of eight Advanced Placement Exams* (Unpublished Statistical Report No. SR-98-03). Princeton, NJ: Educational Testing Service.

Myford, C.M., Marr, D.B. & Linacre, J.M. (1996). *Reader calibration and its potential role in equating for the Test of Written English* (ETS Center for Performance Assessment Report No. MS 95-02). Princeton, NJ: Educational Testing Service.

Myford, C.M. & Mislevy, R.J. (1995). *Monitoring and improving a portfolio assessment system* (ETS Center for Performance Assessment Report No. MS 94-05). Princeton, NJ: Educational Testing Service.

Myford, C.M. & Wolfe, E.W. (2001). *Detecting and measuring rater effects using many-facet Rasch measurement: An instructional module.* Manuscript submitted for publication.

O'Neill, T.R. & Lunz, M.E. (1996, April). *Examining the invariance of rater and project calibrations using a multifacet Rasch model.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

O'Neill, T.R. & Lunz, M.E. (2000). A method to study rater severity across several administrations. In M. Wilson & G. Engelhard, Jr. (Eds.), *Objective Measurement: Theory into Practice* (Vol. 5, pp. 135–146). Stamford, CT: Ablex.

Peterson, S. & Bainbridge, J. (1999). Teachers' gendered expectations and their evaluation of student writing. *Reading Research and Instruction,* 38(3), 255–271.

Pula, J.J. & Huot, B.A. (1993). A model of background influences on holistic raters. In M.M. Williamson & B.A. Huot (Eds.), *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations* (pp. 237–265). Cresskill, NJ: Hampton Press.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Raymond, M.R. (1986). Missing data in evaluation research. *Evaluation and the Health Professions, 9*, 395–420.

Raymond, M.R. & Houston, W.M. (1990, April). *Detecting and correcting for rater effects in performance assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED336429)

Raymond, M.R. & Viswesvaran, C. (1993). Least-squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement,* 30(3), 253–268.

Raymond, M.R. Webb, L.C. & Houston, W.M. (1991). Correcting performance-rating errors in oral examinations. *Evaluation and the Health Professions,* 14(1), 100–122.

Rothenberg, L. Dabney, M. Garner M. & Monsaas, J. (1995, April). *A comparison of methods to estimate the variance-covariance matrix with matrix sampled data.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rudner, L.M. (1992). *Reducing errors due to the use of judges.* ERIC/TM Digest. (Report EDO-TM-92-10). Washington, DC: American Institutes for Research. (ERIC Document Reproduction Service No. ED355254)

Shohamy, E. Gordon, C.M. & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*, 27–33.

Stone, M. & Wright, B.D. (1988). *Separation statistics in Rasch measurement* (Research Memorandum No. 51). Chicago: MESA Press.

Webb, L.C. Raymond, M.R. & Houston, W.M. (1990, April). *Rater stringency and consistency in performance assessment. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.* (ERIC Document Reproduction Service No. ED318776)

Weigle, S.C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing,* 6 (2), 145–178.

Wilson, H.G. (1988). Parameter estimation for peer grading under incomplete design. *Educational and Psychological Measurement, 48*, 69–81.

Wilson, M. & Case, H. (2000). An examination of variation in rater severity over time: A study in rater drift. In M. Wilson & G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 113–133). Stamford, CT: Ablex.

Wolcott, W. et al. (1988). *Discrepancies in essay scoring* (Report No. TM013018). Springfield, VA: TM Clearinghouse. (ERIC Document Reproduction Service No. ED306246)

Wolfe, E.W. Engelhard, G. Jr. & Myford, C.M. (2001, May). *Monitoring reader performance and DRIFT in the AP English Literature and Composition Exam using benchmark essays.* A proposal funded by the Advanced Placement Research and Development Committee, Educational Testing Service, Princeton, NJ.

Wolfe, E.W. & Feltovich, B. (1994, April). *Learning to rate essays: A study of scorer cognition.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED368777)

Wolfe, E.W. & Kao, C. (1996, April). *Expert/novice differences in the focus and procedures used by essay scorers.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY. (ERIC Document Reproduction Service No. ED399286)

Wolfe, E.W. & Kao, C. (1996, April). *The relationship between scoring procedures and focus and the reliability of direct writing assessment scores.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY. (ERIC Document Reproduction Service No. 396005)

Wood, R. & Wilson, D. (1974). Evidence for differential marking discrimination among examiners of English. *The Irish Journal of Education,* 8(1), 36–48.

Wright, B.D. (1984). Despair and hope for educational measurement. *Contemporary Education Review,* 3(1, 281–285.)

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

Wright, B.D. & Stone, M.H. (1979). *Best test design.* Chicago, IL: MESA Press, University of Chicago.

# APPENDIX

**Table 4.2 — Scoring Worksheet — English Literature and Composition**

**Section I: Multiple Choice**

[ _____ (Number correct, out of 55) − (1/4 × _____ (Number wrong)) ] × 1.2272 = _____ (Multiple-Choice Score) = _____ (Weighted Section I Score)

**Section II: Free Response**

Question 1 _____ (out of 9) × 3.0556 = _____ (Do not round)

Question 2 _____ (out of 9) × 3.0556 = _____ (Do not round)

Question 3 _____ (out of 9) × 3.0556 = _____ (Do not round)

Sum = _____ Weighted Section II Score (Do not round)

**Composite Score**

_____ (Weighted Section I Score) + _____ (Weighted Section II Score) = _____ (Round to nearest whole number.)

**AP Grade Conversion Chart**

| Composite Score Range* | AP Grade |
|---|---|
| 108-150 | 5 |
| 94-107 | 4 |
| 75-93 | 3 |
| 47-74 | 2 |
| 0-46 | 1 |

*The candidates' scores are weighted according to formulas determined in advance each year by the Development Committee to yield raw composite scores; the Chief Faculty Consultant is responsible for converting composite scores to the 5-point AP scale.

**Figure A1.** Description of scoring system (The College Board, 1999a, p. 72).

TABLE A1

**Free-Response Question 1 from the 1999 AP English Literature and Composition Exam**[12]

(Suggested time — 40 minutes. This question counts as one-third of the total essay section score.)
Read the following poem carefully, paying particular attention to the physical intensity of the language. Then write a well-organized essay in which you explain how the poet conveys not just a literal description of picking blackberries but a deeper understanding of the whole experience. You may wish to include analysis of such elements as diction, imagery, metaphor, rhyme, rhythm, and form.

**Blackberry-Picking**

Late August, given heavy rain and sun
For a full week, the blackberries would ripen.
At first, just one, a glossy purple clot
Among others, red, green, hard as a knot.
You ate that first one and its flesh was sweet
Like thickened wine: summer's blood was in it
Leaving stains upon the tongue and lust for
Picking. Then red ones inked up and that hunger
Sent us out with milk cans, pea tins, jam pots
Where briars scratched and wet grass bleached our boots.
Round hayfields, cornfields, and potato drills[1]
We trekked and picked until the cans were full,
Until the tinkling bottom had been covered
With green ones, and on top big dark blobs burned
Like a plate of eyes. Our hands were peppered
With thorny pricks, our palms sticky as Bluebeard's.[2]

We hoarded the fresh berries in the byre.[3]
But when the bath was filled we found a fur,
A rat-grey fungus, glutting on our cache.
The juice was stinking too. Once off the bush
The fruit fermented, the sweet flesh would turn sour.
I always felt like crying. It wasn't fair
That all the lovely canfuls smelt of rot.
Each year I hoped they'd keep, knew they would not.

*--Seamus Heaney*



[1] Planted rows

[2] Bluebeard is a character in a fairy tale who murders his wives.

[3] Barn


[12]Free-response Question 1 is taken from the *1999 AP English Literature and Composition Released Exam*, p. 32.

TABLE A2

**Scoring Guidelines for Question 1 from the 1999 AP English Literature and Composition Exam**[13]

| | |
|---|---|
| **9-8:** | These well-conceived and well-ordered essays provide insightful analysis (implicit as well as explicit) of how Heaney creates and conveys his memory of picking blackberries. They appreciate Heaney's physically-intense language for its vivid literal description, but they also understand the meaning of the experience on a profound, metaphoric level. Although the writers of these essays may offer a range of interpretations and/or choose different poetic elements for emphasis, these papers provide convincing readings of the poem and maintain consistent control over the elements of effective composition, including the language unique to the criticism of verse. Their textual references are apt and specific. Though they may not be error-free, they demonstrate the writers' ability to read poetry perceptively and to write with clarity and sophistication. |
| **7-6:** | These essays reflect a sound grasp of Heaney's poem and the power of its language; but they prove less sensitive than the best essays to the poetic ways that Heaney invests literal experience with strong, metaphoric implications. The interpretations of the poem that they provide may falter in some particulars or they may be less thorough or precise in their discussion of how the speaker reveals the experience of "blackberry-picking." Nonetheless, their dependence on paraphrase, if any, will be in the service of analysis. These essays demonstrate the writers' ability to express ideas clearly, but they do not exhibit the same level of mastery, maturity, and/or control as the very best essays. These essays are likely to be briefer, less incisive, and less well-supported than the 9-8 papers. |
| **5:** | These essays are, at best, superficial. They respond to the assigned task yet probably say little beyond the most easily grasped observations. Their analysis of how the experience of blackberry picking is conveyed may be vague, formulaic, or inadequately supported. They may suffer from the cumulative force of many minor misreadings. They tend to rely on paraphrase but nonetheless paraphrase which contains some implicit analysis. Composition skills are at a level sufficient to convey the writers' thoughts, and egregious mechanical errors do not constitute a distraction. These essays are nonetheless not as well-conceived, organized, or developed as upper-half papers. |
| **4-3:** | These lower-half essays reveal an incomplete understanding of the poem and perhaps an insufficient understanding of the prescribed task as well: they may emphasize literal description without discussing the deeper implications of the blackberry-picking experience. The analysis may be partial, unconvincing, or irrelevant—or it may rely essentially on paraphrase. Evidence from the text may be meager or misconstrued. The writing demonstrates uncertain control over the elements of composition, often exhibiting recurrent stylistic flaws and/or inadequate development of ideas. Essays scored 3 may contain significant misreading and/or unusually inept writing. |
| **2-1:** | These essays compound the weaknesses of the papers in the 4-3 range. They may seriously misread the poem. Frequently, they are unacceptably brief. They are poorly written on several counts and may contain many distracting errors in grammar and mechanics. Although some attempt may have been made to respond to the question, the writers' assertions are presented with little clarity, organization, or support from the text of the poem. |
| **0** | A response with no more than a reference to the task. |
| **—** | Indicates a blank response or one that is completely off topic. |



[13] The scoring guidelines for Question 1 are taken from the *1999 AP English Literature and Composition Released Exam*, p. 37-38.

TABLE A3

**Free-Response Question 2 from the 1999 AP English Literature and Composition Exam**[14]

(Suggested time—40 minutes. This question counts as one-third of the total essay section score.)

2. In the following passage from Cormac McCarthy's novel *The Crossing* (1994), the narrator describes a dramatic experience. Read the passage carefully. Then, in a well-organized essay, show how McCarthy's techniques convey the impact of the experience on the main character.

By the time he reached the first talus[1] slides under the tall escarpments[2] of the Pilares the dawn was not far to come. He reined the horse in a grassy swale and stood down and dropped the reins. His trousers were stiff with blood. He cradled the wolf in his arms and lowered her to the ground and unfolded the sheet. She was stiff and cold and her fur was bristly with the blood dried upon it. He walked the horse back to the creek and left it standing to water and scouted the banks for wood with which to make a fire. Coyotes were yapping along the hills to the south and they were calling from the dark shapes of the rimlands above him where their cries seemed to have no origin other than the night itself.

He got the fire going and lifted the wolf from the sheet and took the sheet to the creek and crouched in the dark and washed the blood out of it and brought it back and he cut forked sticks from a mountain hackberry and drove them into the ground with a rock and hung the sheet on a trestlepole where it steamed in the firelight like a burning scrim standing in a wilderness where celebrants of some sacred passion had been carried off by rival sects or perhaps had simply fled in the night at the fear of their own doing. He pulled the blanket about his shoulders and sat shivering in the cold and waiting for the dawn that he could find the place where he would bury the wolf. After a while the horse came up from the creek trailing the wet reins through the leaves and stood at the edge of the fire.

He fell asleep with his hands palm up before him like some dozing penitent. When he woke it was still dark. The fire had died down to a few low flames seething over the coals. He took off his hat and fanned the fire with it and coaxed it back and fed the wood he'd gathered. He looked for the horse but could not see it.

The coyotes were still calling all along the stone ramparts of the Pilares and it was graying faintly in the east. He squatted over the wolf and touched her fur. He touched the cold and perfect teeth. The eye turned to the fire gave back no light and he closed it with his thumb and sat by her and put his hand upon her bloodied forehead and closed his own eyes that he could see her running in the mountains, running in the starlight where the grass was wet and the sun's coming as yet had not undone the rich matrix of creatures passed in the night before her. Deer and hare and dove and groundvole all richly empaneled on the air for her delight, all nations of the possible world ordained by God of which she was one among and not separate from. Where she ran the cries of the coyotes clapped shut as if a door had closed upon them and all was fear and marvel. He took up her stiff head out of the leaves and held it or he reached to hold what can not be held, what already ran among the mountains at once terrible and of a great beauty, like flowers that feed on flesh. What blood and bone are made of but can themselves not make on altar nor by any wound of war. What we may well believe has power to cut and shape and hollow out the dark form of the world surely if wind can, if rain can. But which cannot be held never be held and is no flower but is swift and a huntress and the wind itself is in terror of it and the world cannot lose it.

[1] A sloping mass of rock debris at the base of a cliff

[2] Steep slopes


[14] Free-response Question 2 is taken from the *1999 AP English Literature and Composition Released Exam*, p. 33.

TABLE A4

**Scoring Guidelines for Question 2 from the 1999 AP English Literature and Composition Exam**[15]

| | |
|---|---|
| **9-8:** | The writers of these well-constructed essays define the dramatic nature of the experience described in Cormac McCarthy's passage and ably demonstrate how the author conveys the impact of the experience upon the main character. Having fashioned a convincing thesis about the character's reaction to the death of the wolf, these writers support their assertions by analyzing the use of specific literary techniques (such as point of view, syntax, imagery, or diction) that prove fundamental to their understanding of McCarthy's narrative design. They make appropriate references to the text to illustrate their argument. Although not without flaws, these essays reflect the writers' ability to control a wide range of the elements of effective writing to provide a keen analysis of a literary text. |
| **7-6:** | Developing a sound thesis, these writers discuss with clarity and conviction both the character's response to the death of the wolf and certain techniques used to convey the impact this experience has upon the main character. These essays may not be entirely responsive to the rich suggestiveness of the passage or as precise in describing the dramatic impact of the event. Although they provide specific references to the text, the analysis is less persuasive and perhaps less sophisticated than papers in the 9-8 range: they seem less insightful or less controlled, they develop fewer techniques, or their discussion of details may be more limited. Nonetheless, they confirm the writers' ability to read literary texts with comprehension and to write with organization and control. |
| **5:** | These essays construct a reasonable if reductive thesis; they attempt to link the author's literary techniques to the reader's understanding of the impact of the experience on the main character. However, the discussion may be superficial, pedestrian, and/or lacking in consistent control. The organization may be ineffective or not fully realized. The analysis is less developed, less precise, and less convincing than that of upper half essays; misinterpretations of particular references or illustrations may detract from the overall effect. |
| **4-3:** | These essays attempt to discuss the impact of this dramatic experience upon the main character—and perhaps mention one or more techniques used by McCarthy to effect this end. The discussion, however, may be inaccurate or undeveloped. These writers may misread the passage in an essential way, rely on paraphrase, or provide only limited attention to technique. Illustrations from the text tend to be misconstrued, inexact, or omitted altogether. The writing may be sufficient to convey ideas, although typically it is characterized by weak diction, syntax, grammar, or organization. Essays scored 3 are even less able and may not refer to technique at all. |
| **2-1:** | These essays fail to respond adequately to the question. They may demonstrate confused thinking and/or consistent weaknesses in grammar or another basic element of composition. They are often unacceptably brief. Although the writer may have made some attempt to answer the question, the views presented have little clarity or coherence; significant problems with reading comprehension seem evident. Essays that are especially inexact, vacuous, and/or mechanically unsound should be scored 1. |
| **0** | A response with no more than a reference to the task. |
| **—** | Indicates a blank response or one that is completely off topic. |



[15] The scoring guidelines for Question 2 are taken from the *1999 AP English Literature and Composition Exam*, p. 49-50.

TABLE A5

**Free-Response Question 3 from the 1999 AP English Literature and Composition Exam**[16]

(Suggested time—40 minutes. This question counts as one-third of the total essay section score.)

The eighteenth-century British novelist Laurence Sterne wrote, "No body, but he who has felt it, can conceive what a plaguing thing it is to have a man's mind torn asunder by two projects of equal strength, both obstinately pulling in a contrary direction at the same time."

From a novel or play choose a character (not necessarily the protagonist) whose mind is pulled in conflicting directions by two compelling desires, ambitions, obligations, or influences. Then, in a well-organized essay, identify each of the two conflicting forces and explain how this conflict within one character illuminates the meaning of the work as a whole. You may use one of the novels or plays listed below or another novel or play of similar literary quality.

*The Adventures of Huckleberry Finn*
*Anna Karenina*
*Antigone*
*The Awakening*
*Beloved*
*Billy Budd*
*Ceremony*
*Crime and Punishment*
*Dr. Faustus*
*An Enemy of the People*
*Equus*
*A Farewell to Arms*
*The Glass Menagerie*
*Hamlet*
*Heart of Darkness*
*Jane Eyre*
*Jasmine*
*Light in August*
*A Lesson Before Dying*
*Macbeth*
*The Mayor of Casterbridge*
*Native Speaker*
*The Piano Lesson*
*A Portrait of the Artist as a Young Man*
*A Raisin in the Sun*
*The Scarlet Letter*
*Wuthering Heights*


[16] Free-response Question 3 is taken from in the *1999 AP English Literature and Composition Released Exam,* p. 34

TABLE A6

**Scoring Guidelines for Question 3 from the 1999 AP English Literature and Composition Exam**[17]

| | |
|---|---|
| **9-8:** | Having chosen a novel or play of recognized literary merit, the able writers of these well-ordered essays focus on an appropriate character "whose mind is pulled in conflicting directions by two compelling desires, ambitions, obligations, or influences." By explaining with clarity and precision the nature of the opposing forces with which the character struggles, as well as the implications of this character's internal conflict for the meaning of the work as a whole, these writers manage to construct a compelling argument that illuminates both character and text. Comprehensive in their grasp of their novel or play, these writers neither over-simplify the complex moral dilemmas that often result from the pull of competing forces "of equal strength"; nor do they ignore the ambiguities that make resolution of such conflicts difficult or even impossible. Specific textual references and solid literary analysis support their assertions and demonstrate their own facility with language. |
| **7-6:** | The writers of these essays select both an appropriate text and character, and they provide a clear and coherent discussion of the struggle with opposing forces that goes on within the mind of a character and a persuasive explanation as to how this conflict "illuminates the meaning of the work as a whole." They display sound knowledge of the text, as well as an ability to order ideas and to write with both clarity and creativity. However, the analysis in these essays is less perceptive, less thorough, and/or less specific than the essays above: neither substance nor style is quite so impressive as the 9-8 essays. |
| **5:** | Although these lower-half essays are often characterized by shallow, unsupported generalizations, they provide at least a plausible argument. These writers identify apt characters in well-chosen texts. Their understanding of the concepts prompted by this question may remain inchoate and/or have little to do with literary constructions: instead of focusing on the pull of opposing forces upon the mind of one character, they may discuss a conflict between two or more characters–or another sort of struggle altogether. The attempt to relate the character's conflict to the meaning of the work may be limited or nonexistent. Competent plot summary may substitute for analysis, and references to the text may be limited, random, or vague. The writing in these essays does not usually demonstrate consistent control over the elements of composition. |
| **4-3:** | These lower-half papers convey a less than adequate comprehension of the assignment. They choose a more or less appropriate text, and they make a reasonable selection of a character from that text. Their discussion of conflicting forces will undoubtedly falter, however, and they may do little to explore the implications of the character's struggle for the meaning of the work as a whole. They seldom exhibit compelling authority over the selected text. Though these essays offer at least a rudimentary argument, support usually depends on unsubstantiated generalizations rather than specific examples. These essays may contain significant misinterpretations and displace analysis with paraphrase or plot summary. The writing may be sufficient to convey some semblance of the writer's ideas, but it reveals only limited control over diction, organization, syntax, or grammar. |
| **2-1:** | These essays compound the weakness of essays in the 4-3 range. They may seriously misread the novel or the play, or the question itself. They may choose a problematic work. They may contain little, if any, clear, coherent argument: they provide impressions rather than analysis. In addition, they are poorly written on several counts, including many distracting errors in grammar and mechanics, or they are unacceptably brief. Essays that are especially vacuous, ill-organized, illogically argued, and/or mechanically unsound should be scored 1. |
| **0** | A response with no more than a reference to the task. |
| **—** | Indicates a blank response or one that is completely off topic. |



[17] The scoring guidelines for Question 3 are taken from in the *1999 AP English Literature and Composition Released Exam,* erratum notice.