Performance of Men and Women on Multiple-Choice and Constructed-Response Tests for Beginning Teachers

Samuel A. Livingston

Stacie L. Rupp

Performance of Men and Women on Multiple-Choice and Constructed-Response Tests for Beginning Teachers

Samuel A. Livingston and Stacie L. Rupp

ETS, Princeton, NJ

November 2004

Abstract

Some previous research results imply that women tend to perform better, relative to men, on constructed-response (CR) tests than on multiple-choice (MC) tests in the same subjects. An analysis of data from several tests used in the licensing of beginning teachers supported this hypothesis, to varying degrees, in most of the tests investigated. The hypothesis was strongly supported in Praxis™ Principles of Learning and Teaching tests for secondary school teachers and in subject-knowledge tests for social studies teachers, science teachers, and middle school mathematics teachers. Evidence for the hypothesis was weak in subject-knowledge tests for middle school English teachers and for secondary school mathematics teachers. Subject-knowledge tests for secondary school English teachers did not show the hypothesized relationship. The analysis was based on plots showing the cumulative percentages of men and women attaining each possible score on the MC and CR tests.

Key words: Gender differences, multiple-choice, constructed-response, teacher licensing, p-p plot

The study described in this paper was a comparison of the performance of men and women on multiple-choice (MC) and constructed-response (CR) tests in the same subject fields. The purpose of the study was to investigate the hypothesis that women tend to perform better, in comparison to men, on CR tests than on MC tests that measure closely related knowledge and skills.

Some previous research studies involving high school students have compared the performance of male and female test takers on MC and CR tests in the same subject. The results have generally shown that, compared with the male students, the female students performed better on the CR tests than on the MC tests. A study of performance on the College Board's English Composition Test with Essay (Petersen & Livingston, 1982) found such a difference when comparing male and female students within each of four ethnic subgroups. Subsequent large-scale studies involving Advanced Placement Program® (AP®) examinations (Mazzeo, Schmitt, & Bleistein, 1992; Willingham & Cole, 1997) showed that the strength of this effect varied across academic subjects. The effect was largest in the social sciences; smaller but substantial in English and the natural sciences; and near zero in mathematics, computer science, and foreign languages. Examinations taken by secondary school students in England have shown results that followed the same pattern, though not as consistently, according to a study by Murphy in 1982 (cited in Willingham & Cole, 1997). Breland, Danos, Kahn, Kubota, and Bonner (1994) conducted a study involving a more detailed scoring of responses to the AP exam in U.S. history and concluded that "the most reasonable explanation of these format gender differences is that the two types of tests [MC and CR] measure different skills, both of which are important in history."

The present study involved a different population of test takers—beginning teachers—and a different set of tests—the Praxis II™ Subject Assessments. To make the kind of comparisons that are the focus of this study, it is necessary to have both MC and CR scores from a group that includes fairly large numbers of both men and women. Data from these tests offered the opportunity to make three kinds of male/female performance comparisons:

- performance on MC and CR questions based on the same stimulus material

- performance on separate MC and CR sections of the same test

- performance on separate MC and CR tests in the same academic subject

1

These three comparisons are based, respectively, on data from

- Praxis Principles of Learning and Teaching (PLT) tests for teachers of grades K–6 and for teachers of grades 7–12

- tests of subject knowledge in English, social studies, science, and mathematics for middle school teachers in each of these subjects

- tests of subject knowledge in English, social studies, biology, and mathematics for secondary school teachers in each of these subjects

The selection of tests for the study was intended to represent a broad sampling of the kinds of content in which beginning teachers are tested as a condition for licensure. Several features of the tests selected for the study are summarized in Appendix A.

The analyses in this study required data from men and women who had taken both the MC and CR questions. On the tests of subject knowledge for secondary school teachers, the MC and CR questions were contained in separate tests, and the scores on each test (MC and CR) were scaled for comparability across forms. On these tests, it was necessary to restrict the analysis to test takers who took both the MC and CR tests, but it was possible to make comparisons based on the combined data from three forms of each test. On the other tests, the MC and CR questions were contained in the same test, and the MC and CR scores were raw scores. On those tests, it was necessary to analyze the data from each form separately, but all the test takers who took a particular form could be included in the analysis for that form.

Method

The analysis technique used in this paper has three essential characteristics:

- It treats the two groups (male and female test takers) and the two types of test (MC and CR) in the same way. It does not use one group or one type of test as a standard and compare the other group or the other type of test against it.

- It makes performance comparisons over the full range of proficiency represented by the test takers included in the analysis. It is not limited to a comparison of the means and standard deviations of the scores, or of a few selected percentiles of the score distributions.

- It presents the male/female performance comparisons graphically, making it possible to see at a glance the way in which the performance of men and women differed on the two types of tests. It also includes a summary statistic that summarizes this comparison in a single number.

The analyses are based on the percentages of male and female test takers attaining each possible score or a higher score (i.e., the cumulative percentage, accumulating from high to low). Each analysis consists of a graph with two curves, one for MC scores and one for CR scores, as shown in Figure 1. Each curve consists of a series of connected data points, corresponding to the possible scores on the test. The horizontal position of the data point—its x value—indicates the percentage of the women attaining *at least* that score; the vertical position of the point—its y value—indicates the percentage of the men attaining *at least* that score. Each graph also includes a dotted line representing the identity line ($y = x$). If the percentage of test takers attaining at least a particular score is the same for women as for men, the data point for that score will lie on the identity line. If the percent attaining at least that score is higher for men than for women, the data point will lie above and to the left of the identity line. If the percent attaining at least that score is higher for women than for men, the data point will lie below and to the right of the identity line. The greater the male/female difference, the farther the data point will depart from the identity line.[1]

This approach makes it possible to compare the relative performance of men and women on MC and CR tests, *even if the scores on the MC test and the scores on the CR test are not comparable.* The numbers being compared are not scores, but percentages of test -takers. For example, in Figure 1, the women outperformed the men at all score levels on the MC portion of the test. The percentage of the women attaining any given MC score was higher than the percentage of the men attaining that MC score. Therefore, the x coordinate is greater than the y coordinate for each data point on the curve for the MC portion of the test, and the curve for the MC portion of the test lies entirely to the right of the identity line. But on the CR portion of the test, the women outperformed the men by an even larger margin. Therefore, the data points that define the curve for the CR portion of the test are even farther to the right of the identity line than those defining the curve for the MC portion; and the curve for the CR portion lies entirely to the right of the curve for the MC portion.
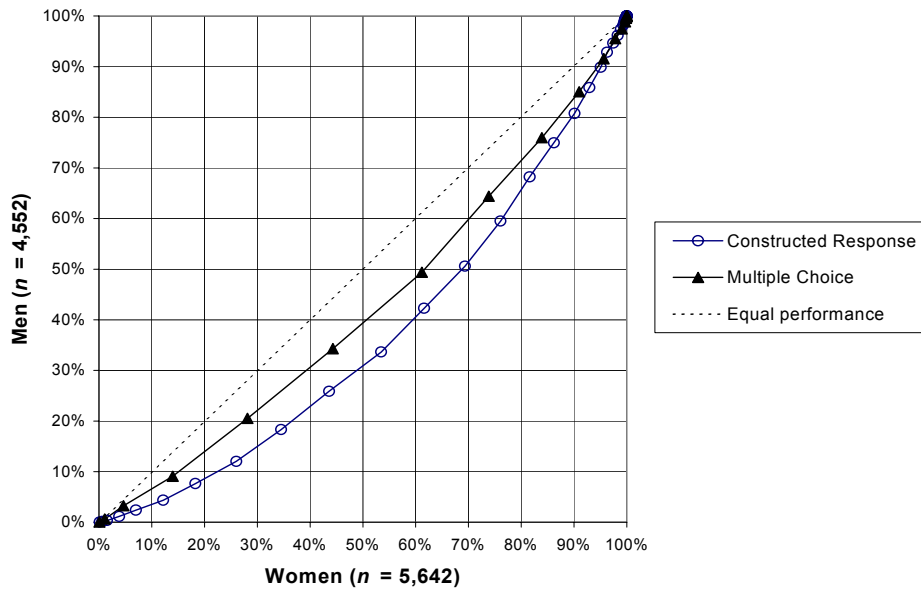
Figure 1. **Results for one form of PLT 7–12 test.**

Notice that the data points in the upper right corner represent the scores for which the percentage of test takers attaining at least those scores is very high (i.e., the lowest scores). Conversely, the data points in the lower left corner represent the highest scores (those attained by only a small percentage of the test takers).

The curves in Figure 1 are quite smooth because the score distributions are smooth in both groups of test takers, as might be expected with groups of this size (more than 4,000 in each group). When one of the groups of test takers is not large, the score distributions tend not to be smooth, and the curves in the plot are not smooth. Figure 2 shows a graph in which the smaller group included only 97 men (and the larger group only 207 women). The curves are not smooth. Nevertheless, the results indicated by the graph are clear. The curve for the MC test lies above the identity line, indicating that the men outperformed the women. The curve for the CR test lies quite close to the identity line, indicating that the men and the women performed about equally well.

To summarize the difference between the MC and CR results in a single number, we computed the *signed area* between the curves, expressed as a percentage of the total area in the graph. We decided, arbitrarily, to refer to an area between the curves as *positive* if the curve for the CR test is below and to the right of the curve for the MC test, as in Figure 1. If the curve for the MC test is below and to the right of the curve for the CR test, we refer to the area as *negative*.

If the curves cross, the negative area is subtracted from the positive area. Thus, a positive value for the signed area indicates a test on which the women performed better, relative to the men, on the CR test rather than on the MC test. In Figure 1, the signed area between the curves is 5.7%. In Figure 2, the signed area between the curves is 7.2%.
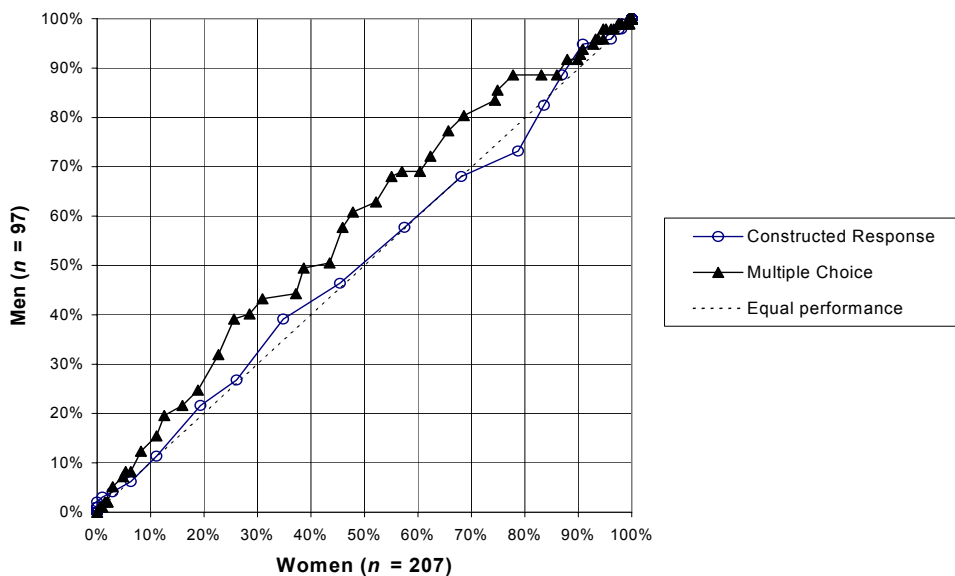
*Figure 2.* **Results for a test taken by a smaller number of test takers.**

Results

Principles of Learning and Teaching Tests

The first set of male/female performance comparisons in this study are based on PLT tests administered before September 2002. Each of these tests included MC and CR questions based on the *same stimulus material*. Each test form included three *cases*. A case consisted of a written description of a teaching situation (about two printed pages), followed by seven multiple-choice questions and two constructed-response questions related to the situation described.[2] Each test form also included several multiple-choice questions not linked to any case; those questions were not included in the analysis for this study. Each CR question called for the test taker to write a response of about five sentences. There are PLT tests for teachers of grades K–6, grades 5–9, and grades 7–12. (A PLT Early Childhood test has recently been added.) We analyzed data from five forms of the PLT K–6 test and five forms of the PLT 7–12 test. The scores in our

analysis are raw scores and are not comparable across forms of the test; therefore, the MC/CR comparisons must be made separately for each form. For each test form that we investigated, we accumulated data over two-or-more test administrations.

Figure 1 shows the results for a typical form of the PLT test for grades 7–12. Because of the large numbers of test takers, the curves are quite smooth. On all five forms that we investigated, the women outperformed the men on both the MC and CR questions. And on all five forms, the women outperformed the men by a greater margin on the CR questions than on the MC questions that were based on the same stimulus material. However, the size of this difference varied from one form to another. The signed area statistics for the five forms were 8.7%, 7.4%, 5.7%, 5.5%, and 3.5%.

Figure 3 shows the results for one form of the PLT K–6 test. Again, the numbers of test takers were large, and the curves were quite smooth. On all five forms that we investigated, the women outperformed the men on both the CR questions and the MC questions based on the same stimulus material, and by a greater margin on the CR questions than on the MC questions. However, the differences between the results for the CR questions and those for the MC questions tended to be smaller than those on the PLT grades 7–12 test. The signed area statistics for the five forms were 6.6%, 4.3%, 3.2%, 2.8%, and 2.0%.
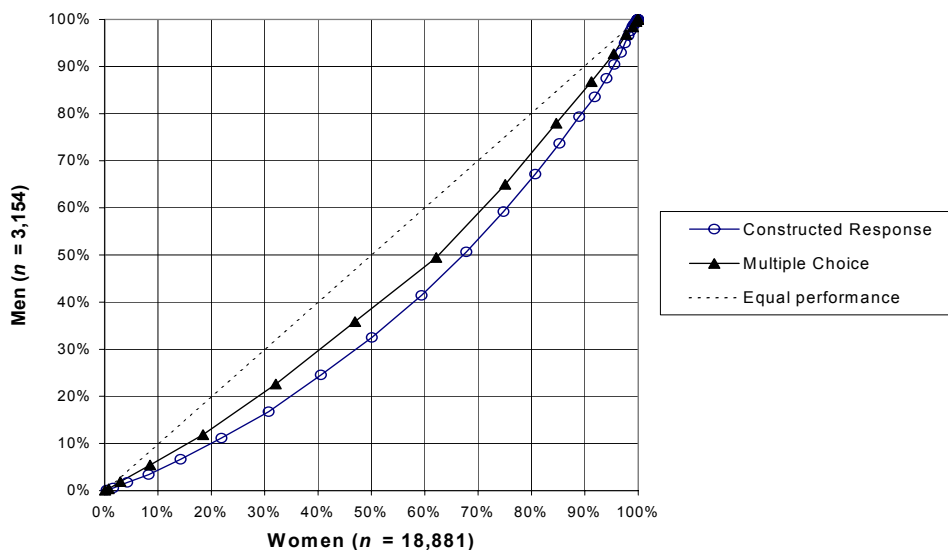
Figure 3. **Results for one form of PLT K–6 test.**

Tests in Academic Subject Areas

We investigated the performance of men and women on tests in four academic subject areas: English, social studies, science, and mathematics. In each subject area, we investigated tests for middle school teachers and tests for secondary school teachers.

In the tests for middle school teachers, the MC and CR questions are contained in the same test (though not based on the same specific content, as were the MC and CR questions in the Principles of Learning and Teaching tests). The MC and CR sections are not separately timed, but the instructions to the test-taker recommend allowing 90 minutes for the MC questions and 30 minutes for the CR questions. The MC and CR section scores are raw scores and are not comparable across forms of the test. We investigated three forms of each test, analyzing the data separately for each form.

In the tests for secondary school teachers, the MC and CR questions are contained in entirely separate tests, but many states require beginning teachers in the subject to take both of those tests. Our analyses were based on data from test takers who took both tests on the same day. In each subject, we investigated three pairs of test forms. We first analyzed the data separately for each pair of forms. We then performed a combined analysis, using scaled scores (which are equated for comparability across forms).

English language arts. The middle school English language arts test is a two-hour test containing 90 MC questions and two CR questions. Each CR question calls for a short essay— about two or three short paragraphs. The MC questions test comprehension and interpretation of literary texts and knowledge of terms and concepts. The two CR questions test literary analysis and rhetorical analysis. Each CR question presents a short written selection and asks the test taker to explain some feature of the author's use of language. The women tended to outperform the men on both the MC and the CR portions of the test, but the male/female differences were much smaller than those observed with the PLT tests. The results were not strongly consistent from form to form. On two of the three forms, the male/female difference (in favor of the women) was larger on the CR section than on the MC section. On the remaining form, the difference (in favor of the women) was larger on the MC section than on the CR section. The signed area statistics for the three forms were 3.8%, 1.6%, and -1.8%.

Figure 4 shows the results for one of the three forms. On this form, the comparison between the MC and CR results was somewhat complex. The women in the middle of the score

distribution did better, in comparison to the men, on the CR section than on the MC section. However, the low-scoring women did better, in comparison to the men, on the MC section than on the CR section. And the highest-scoring women, in comparison to the men, did about equally well on the MC and CR sections.
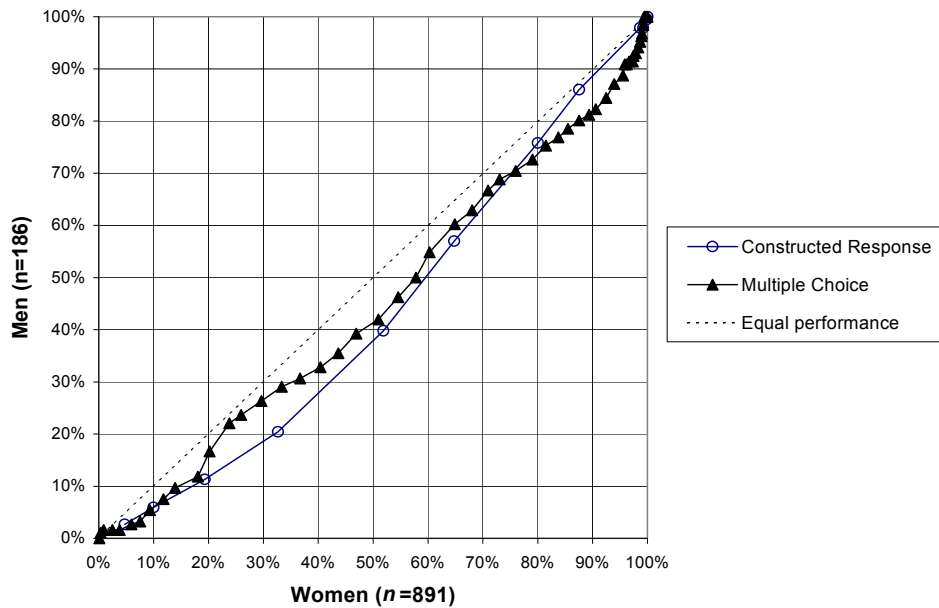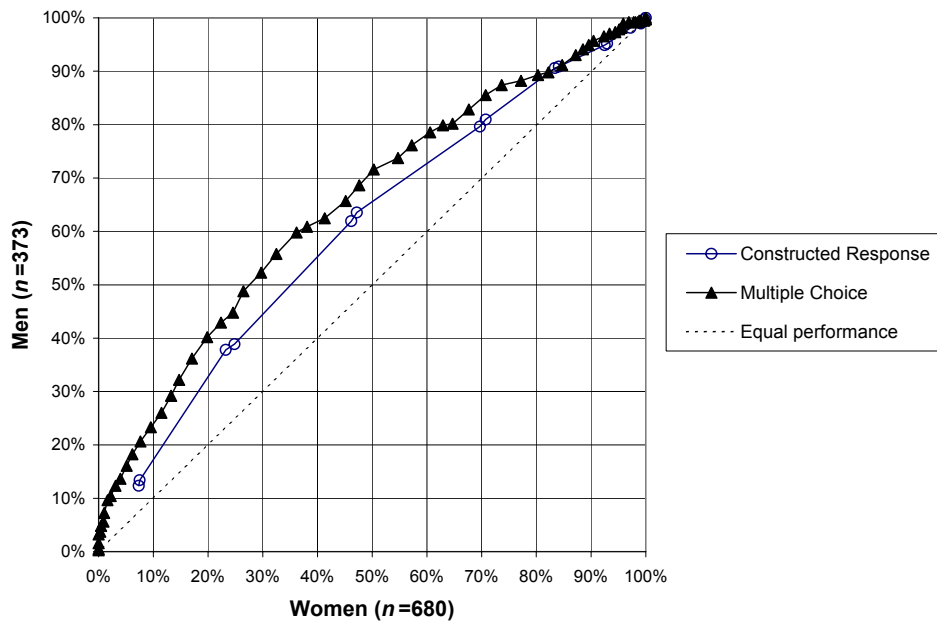

Figure 4. **Results for one form of middle school English test.**

The secondary school tests in English language and literature are each two hours long. The MC test consists of 120 questions that test comprehension of literary texts, knowledge of works of literature, and knowledge of terminology. The CR test consists of four essays, two on interpreting literature (one on poetry, one on prose) and two on literary issues (e.g., the role of the writer in society). The male/female differences varied somewhat over the three pairs of test forms that we investigated, but, in general, the male/female differences were small on both the MC tests and the CR tests. The MC and CR results were similar on all three pairs of test forms; the signed area statistics were 1.4%, -0.3%, and -1.7%. In the analysis based on the combined data, shown in Figure 5, the curves for the MC and CR tests were almost indistinguishable. The signed area statistic was -0.1%.
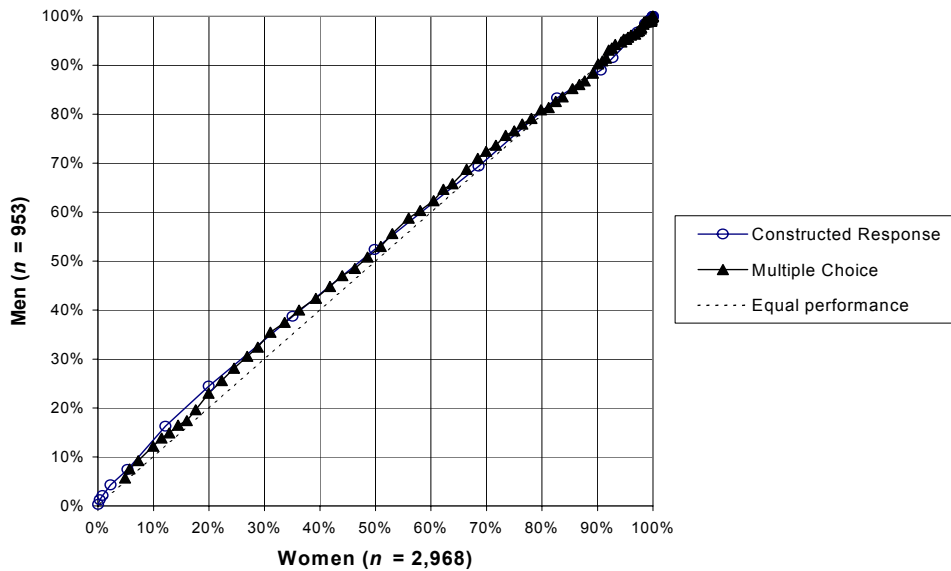
Figure 5. **Results for secondary school English tests, combining data across three forms of each test.**

Social studies. The middle school social studies test is a two-hour test containing 90 MC questions and three CR questions. The MC questions test knowledge of facts, knowledge of terms and concepts, and the application of principles of geography, economics, and so on. Almost half the MC questions test knowledge of history; most of the others test knowledge of civics and government, geography, and economics. The three CR questions each call for a response of two short paragraphs, emphasizing cause/effect relationships. Each question tests knowledge of history and one other area: government, geography, or economics. On all three forms of this test, men clearly outperformed women on both the MC and CR sections of the test. In every case, the male/female difference was greater on the MC section than on the CR section. Figure 6 shows the results for one of the three forms. The signed area statistics for the three forms were 8.3%, 5.0%, and 4.8%.
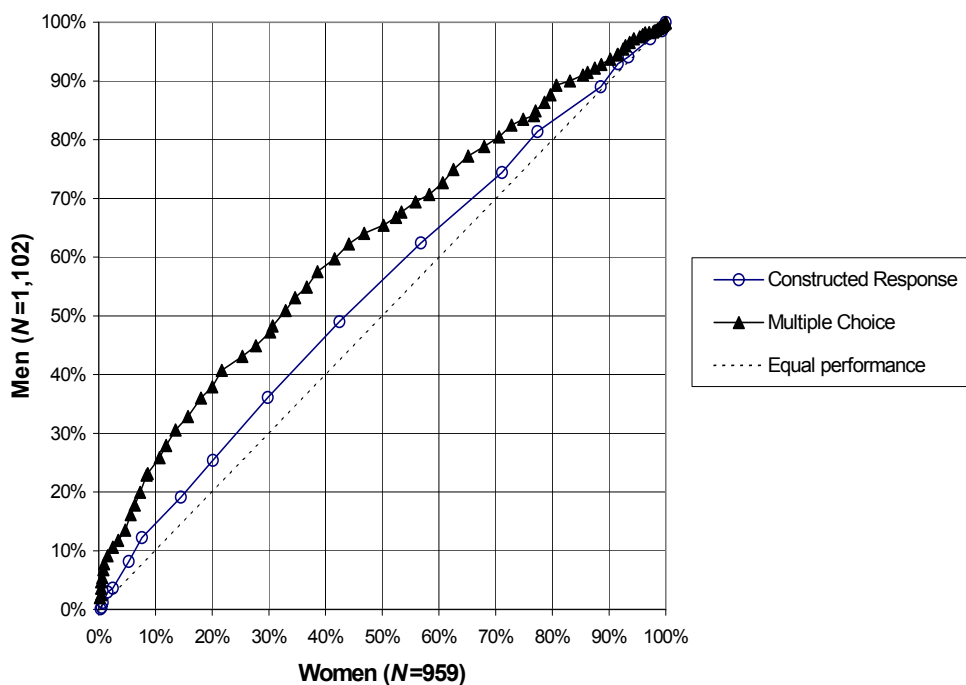
Figure 6. **Results for one form of middle school social studies test.**

The secondary school tests in social studies include a two-hour MC test and a one-hour CR test. The MC test consists of 130 questions, nearly one-half of which are in history. Most of the questions test knowledge of facts or of terms and concepts. The CR test consists of two essay questions, each calling for a response of about four paragraphs. One question requires the test taker to present an argument for or against a position expressed in a quote, a cartoon, or some other brief source document. The other question calls for an analysis of a historical situation.

The men outperformed the women by a large margin on all three forms of the MC test, but by a much smaller margin on the CR test. On two of the three pairs of forms that we investigated, the scores of the men and women on the CR test were nearly equal. The signed area statistics for the three pairs of forms were 10.4%, 7.6%, and 7.0%. Figure 7 shows the results of the analysis based on the combined data. The signed area statistic for this analysis was 8.1%.

10

Figure 7. **Results for secondary school social studies tests, combining data across three forms of each test.**

Science. The middle school science test is a two-hour test containing 90 MC questions and three CR questions. Although many of the MC questions test knowledge of facts, many others require the application of concepts or principles of science. The three CR questions—one from the physical sciences, one from the life sciences, and one from the earth sciences —each call for a response of one or two paragraphs. Typically, they require the test taker to predict or explain observable events or phenomena by applying scientific principles. On all three forms investigated, the men outperformed the women by a wide margin on the MC portion, but by a much narrower margin on the CR portion. Figure 8 shows the results for one of the three forms. The signed area statistics for the three forms were 10.6%, 8.8%, and 6.3%.
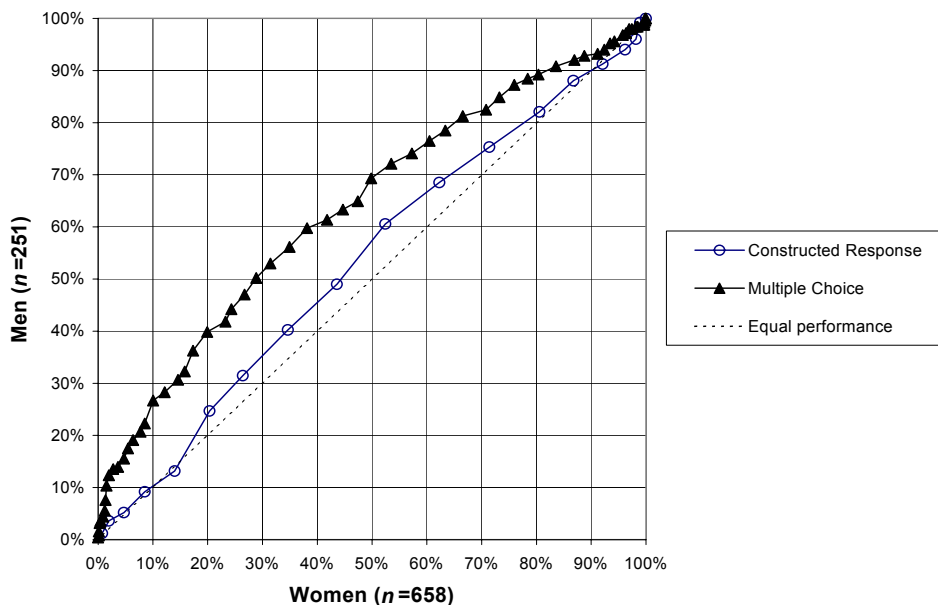
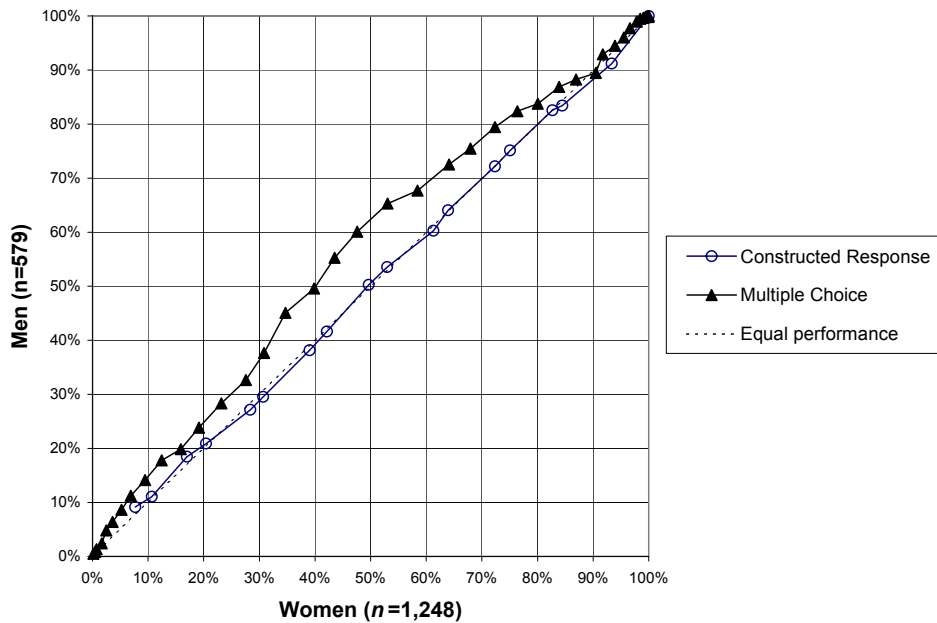*Figure 8.* **Results for one form of middle school science test.**

The secondary school tests in biology include a one-hour MC test and a one-hour CR test. The MC test consists of 75 questions that test knowledge of facts, knowledge of terms and concepts, and the application of principles of the science. The CR test consists of three essay questions that test the test taker's knowledge of important relationships and ability to explain them (e.g., How do principles of genetics account for heritable variation within a population?).

In the separate analyses for the three pairs of forms, the curves were not smooth, and the male/female differences were not consistent across forms, presumably because of the small numbers of test takers. Nevertheless, on all three pairs of forms, the women performed better, compared with the men, on the CR test than on the MC test. The signed area statistics for the three pairs of forms were 8.9%, 7.2%, and 5.6%. In the analysis based on the combined data, the men outperformed the women by a substantial margin on the MC test, but by only a small margin on the CR test. Among the highest-scoring test takers, the women narrowly outperformed the men on the CR test. Figure 9 shows the results of the comparison based on the combined data. The signed area statistic for this analysis was 7.9%.
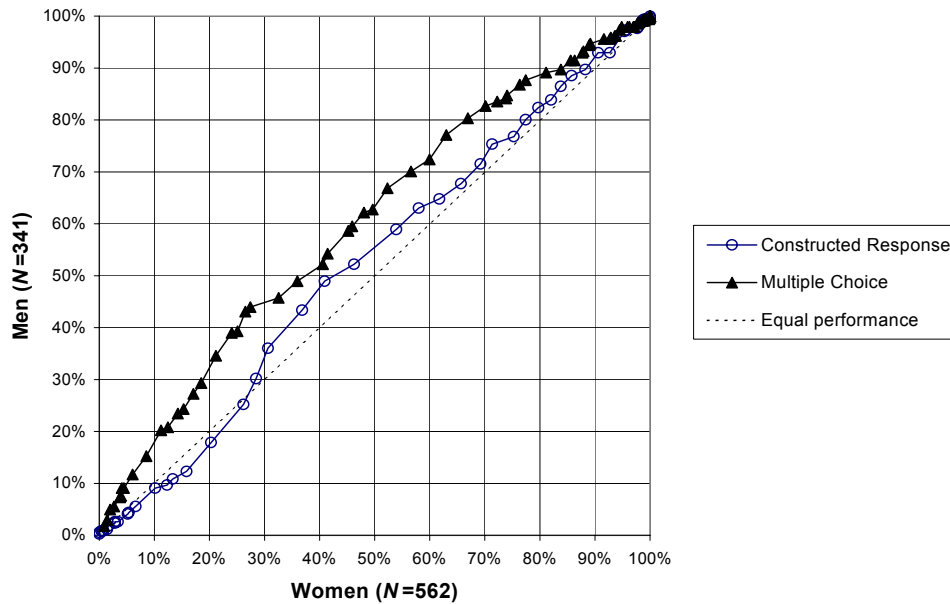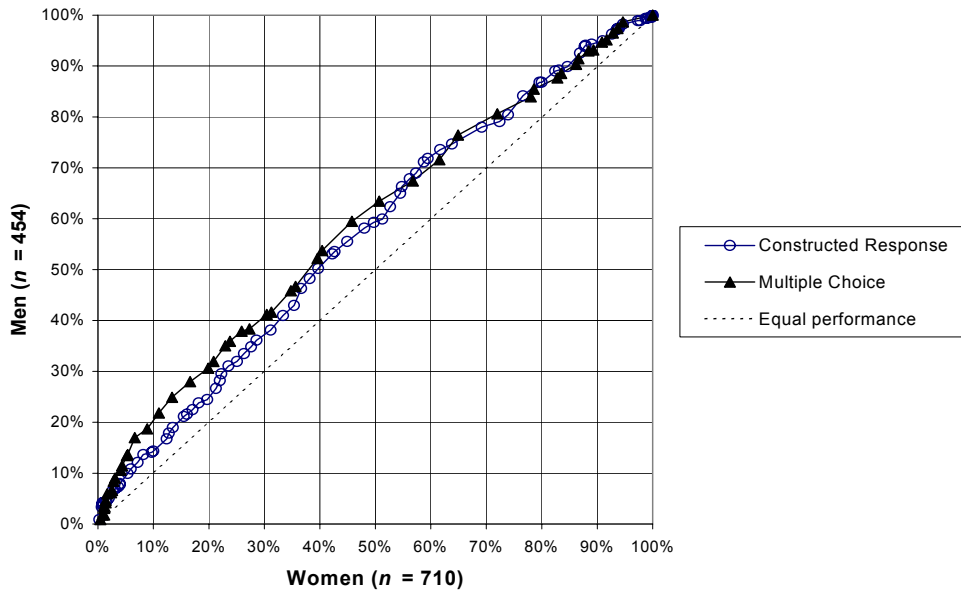
Figure 9. **Results for secondary school biology tests, combining data across three forms of each test.**

Mathematics. The middle school mathematics test is a two-hour test containing 45 MC questions and three CR questions. The MC questions require the test taker to solve problems, interpret graphs and tables, and apply mathematical concepts. The CR questions vary. The question may require the test taker to state and illustrate a definition or to solve a problem (or two or three related short problems) and to explain the reasoning behind the solution. Typically, the test takers communicate their reasoning partly in words and partly in drawings, numbers, or mathematical notation. The men outperformed the women on the MC sections of all three forms, but on two of the three forms, the women performed at least as well as the men on the CR section. Figure 10 shows the results for one of the three forms. Although the male/female differences were not consistent across forms, the difference between the MC and CR results was consistent. The signed area statistics for the three forms were 6.4%, 6.2%, and 5.1%.

Figure 10. **Results for one form of middle school mathematics test.**

The secondary school tests in mathematics include a two-hour MC test and a one-hour CR test. The MC test consists of 50 questions. Most of the MC questions call for the test taker to solve problems, but some require the test taker to apply a mathematical concept without actually solving a problem. The CR test consists of four exercises: one proof, one modeling exercise, and two problems. The CR test requires the test taker to communicate clearly the reasoning behind each solution, using a combination of words, diagrams, and mathematical notation. In the separate analyses for the three pairs of forms, the curves were not smooth and the results differed somewhat across forms, presumably because of the small numbers of test takers. The signed area statistics for the three pairs of forms were 3.6%, 2.6%, and -0.2%. In the analysis based on the combined data, the men outperformed the women on both the MC test and the CR test. The male/female difference was slightly greater on the MC test than on the CR test in the upper half of the score distribution, but about the same on both tests in the lower half of the score distribution. Overall, the difference between the MC and CR results was small. Figure 11 shows the results of the comparison based on the combined data. The signed area statistic for this analysis was 1.9%.

Figure 11. **Results for secondary school mathematics tests, combining data across three forms of each test.**

Summary

Figure 12 summarizes the results of these analyses for the different subjects tested, by showing the signed area statistic for each type of test in each subject. For the secondary school subjects, this statistic is based on the analysis that combined the data from the three pairs of forms investigated. For the other subjects, it is the average value, averaging over the individual test forms investigated. In all the subjects but one (secondary school English), the signed area statistic is positive, indicating that the women performed better, compared with the men, on the CR questions than on the MC questions. The difference between the MC results and the CR results was large for the science and social studies tests, the middle school mathematics test, and the test of Principles of Learning and Teaching for grades 7–12 (but smaller for grades K–6). The difference between the MC and CR results was small for the secondary school math tests, even smaller for the middle school English test, and essentially zero for the secondary school English tests. The results can be summarized verbally as follows:

Women, compared with men, performed *clearly better on CR questions than on MC questions* in

- Principles of Learning and Teaching (especially the secondary school test)

- Social studies

- Science

- Middle school mathematics

Women, compared with men, performed *slightly better on CR questions than on MC questions* in

- Secondary school mathematics

- Middle school English

Women, compared with men, performed *equally well on CR questions and MC questions* in

- Secondary school English
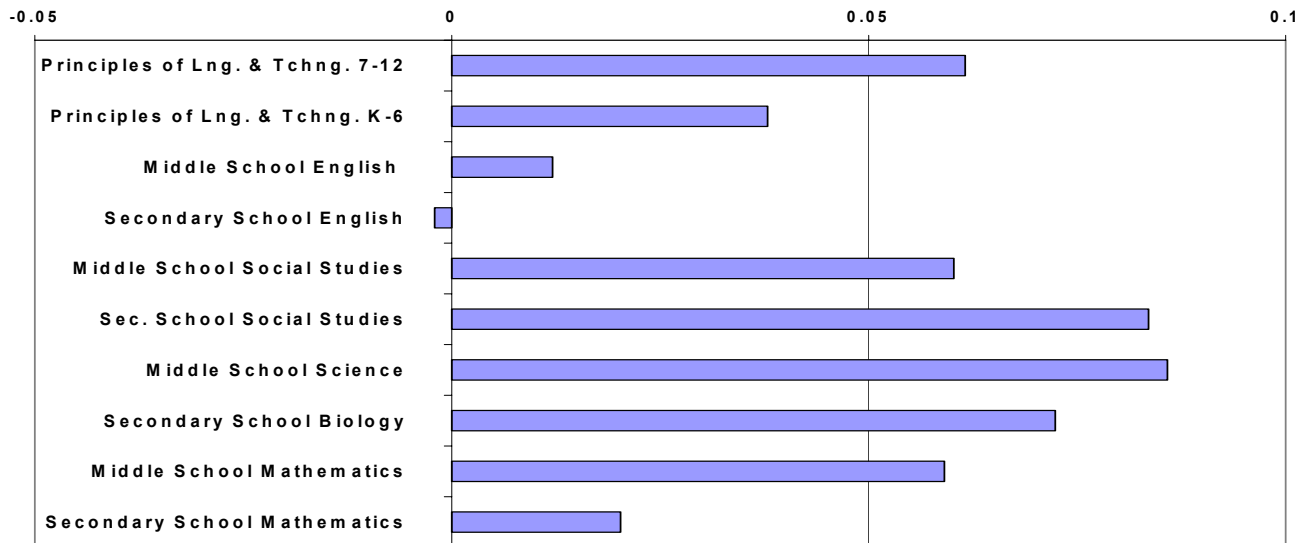
Figure 12. **Signed area statistic, averaging over forms of each test.**

Discussion

Previous research had led us to expect that we would find women performing substantially better, compared with men, on the CR questions than on the MC questions, except possibly on the mathematics tests. In general, the results of the study fit that pattern, but with some notable exceptions, particularly the subject area tests for English teachers. With the exception of those tests, the results might be attributed to male/female differences in the ability to communicate in written language. That skill is not used at all on the MC tests, and it is certainly necessary to do well on the CR tests (though possibly less so in mathematics than in the other subjects). This explanation cannot account for all of the results that we observed—in particular, those for the secondary school English tests. However, it may possibly account for the difference between the results for the middle school and secondary school mathematics tests.

The difference between the results for the middle school and secondary school math tests may have had to do with the difficulty of the CR problems. On the CR test for secondary school math teachers, the main source of difficulty may be in solving the problems. On the CR section of the middle school math test, the main source of difficulty may be in communicating the solution, rather than in solving the problem. If these speculations are correct, and if the women taking these tests are relatively stronger (compared to the men) in communication than in problem solving, we could expect to find results like those we observed.

On the English tests, the differences between the performance of men and women were small and nearly identical for the two types of test (MC and CR). This result contrasts markedly with the results for the science and social studies tests. However, it is the sort of result that could be expected if men who become English teachers write as well as their female counterparts, and if men who become science or social studies teachers do not write as well as their female counterparts.

This hypothesis suggested an additional analysis. The test takers taking the PLT tests had been asked to indicate their major field of study on their registration forms. That information made it possible to reanalyze the data for the PLT 7–12 test, restricting the analysis to those test takers who indicated on the registration form that their college major was in English or in English education. If male/female differences in writing skill are the source of the performance difference on the CR portion of the PLT 7–12 test, and if men who become English teachers

17

write as well as women who become English teachers, the CR curve should resemble the MC curve when the analysis is restricted to English teachers.

The results of this analysis did not support the hypothesis. Figure 13 shows the results of this analysis for the same form of the test as Figure 1. The area between the curves *increased* when the analysis was restricted to English and English Education majors, not only on this form, but on all five forms of the test that we investigated! The signed difference statistics for the five forms were 11.6%, 11.6%, 8.4%, 7.6%, and 5.4%. Averaging across the five forms, the signed area statistic was 8.9% in this analysis, as compared with 6.1% in the analysis that included all test takers.
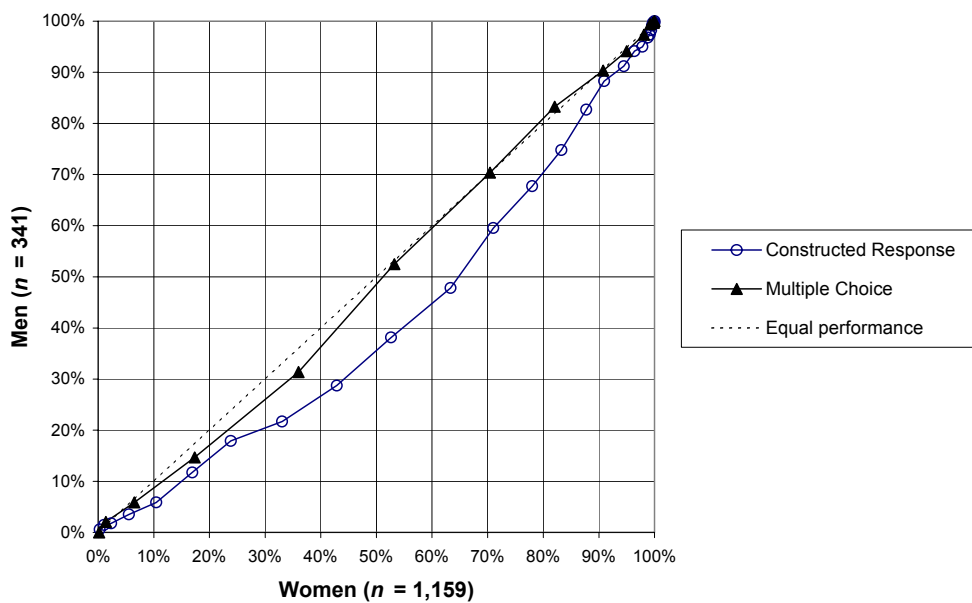
Figure 13. **Results for English majors taking one form of PLT 7–12 test.**

Implications of the Study

It would be wrong to regard the results of this study as establishing general truths about MC and CR test formats. Our study was not a controlled laboratory study; it was a real-world observational study. The test takers were not representative samples of all men and all women; they were beginning teachers seeking licensure in states that required the particular tests they were taking. The comparisons were based entirely on tests from the Praxis Series.

The test developers in the Praxis program do not use a CR test to measure skills that they can measure effectively with a less expensive, more reliable MC test. They use CR tests to measure those skills that they cannot measure effectively with MC tests. In the Praxis program (and, presumably, in other testing applications), the decision to use MC tests, CR tests, or both types of tests is not simply a choice of response formats; it is a choice of which skills to measure. The results of our study imply that this choice, more often than not, will have an effect on the relative performance of men and women taking the test. In many academic subject areas, an economically motivated decision to use only multiple-choice questions can be expected to result in relatively lower performance by women taking the test.

References

Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Bonner, M. W. (1994). Performance versus objective testing and gender: An exploratory study of an Advanced Placement history examination. *Journal of Educational Measurement, 31*, 275–293.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1992). *Sex-related performance differences on constructed-response and multiple-choice sections of Advanced Placement examinations* (College Board Research Rep. No. 92-7; ETS RR-93-05). New York: College Entrance Examination Board.

Petersen, N. S., & Livingston, S. A. (1982). *English composition test with essay: A descriptive study of the relationship between essay and objective scores by ethnic group and sex* (SR-82-96). Princeton, NJ: ETS.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Erlbaum.

Notes

[1] Each of the curves in the graph is similar to what statisticians call a *p-p plot.* Therefore, these graphs could be referred to as *double p-p plots*.

[2] Forms of this test administered since September 2002 follow a somewhat different format; all questions based on the cases are constructed-response.

Appendix A

Characteristics of Multiple-Choice and Constructed-Response Tests in Each Subject

Principles of Learning and Teaching

- Multiple-choice (MC) and constructed-response (CR) questions in same test: 2 hours

- 21 MC questions based on same stimulus material as CR questions

- 6 CR questions: 36 possible points

- CR response: written, approximately 5 sentences

Middle School English

- MC and CR questions in same test: 2 hours (30 minutes recommended for CR)

- 90 MC questions

- 2 CR questions: 12 possible points

- CR response: short essay

Secondary School English

- Separate MC and CR tests: 2 hours each

- 120 MC questions

- 4 CR questions: 24 possible points

- CR response: essay

Middle School Social Studies

- MC and CR questions in same test: 2 hours (30 minutes recommended for CR)

- 90 MC questions

- 3 CR questions: 18 possible points

- CR response: two short paragraphs

Secondary School Social Studies

- Separate 2-hour MC and 1-hour CR tests

- 130 MC questions

- 2 CR questions: 20 possible points

- CR response: essay

Middle School Science

- MC and CR questions in same test: 2 hours (30 minutes recommended for CR)

- 90 MC questions

- 3 CR questions: 18 possible points

- CR response: one or two paragraphs

Secondary School Biology

- Separate MC and CR tests: 1 hour each

- 75 MC questions

- 3 CR questions: 30 possible points

- CR response: essay

Middle School Mathematics

- MC and CR questions in same test: 2 hours (30 minutes recommended for CR)

- 45 MC questions

- 3 CR questions: 18 possible points

- CR response: definitions, hand-drawn graphs, explanation of problem solution

Secondary School Mathematics

- Separate 2-hour MC and 1-hour CR tests

- 50 MC questions

- 4 CR exercises (unequally weighted): 60 possible points

- CR response: present and explain proof, solve problem and explain solution

Appendix B

Signed Area Statistic for Each Test Form (or Pair of Forms) Investigated

Subject	Test form included in study	Signed area between curves (percent of total area of graph)
Principles of Learning and Teaching grades 7–12	1st	5.7%
	2nd	5.5%
	3rd	7.4%
	4th	8.7%
	5th	3.5%
Principles of Learning and Teaching grades K–6	1st	3.2%
	2nd	2.0%
	3rd	2.8%
	4th	4.3%
	5th	6.6%
Middle school English	1st	3.8%
	2nd	1.6%
	3rd	-1.8%
Secondary school English	1st	1.4%
	2nd	-1.7%
	3rd	-0.3%
Middle school social studies	1st	5.0%
	2nd	8.3%
	3rd	4.8%
Secondary school social studies	1st	7.0%
	2nd	7.6%
	3rd	10.4%

(Table continues)

Table (continued)

Subject	Test form included in study	Signed area between curves (percent of total area of graph)
Middle school science	1st	8.8%
	2nd	10.6%
	3rd	6.3%
Secondary school biology	1st	7.2%
	2nd	8.9%
	3rd	5.6%
Middle school mathematics	1st	5.1%
	2nd	6.4%
	3rd	6.2%
Secondary school mathematics	1st	3.6%
	2nd	2.6%
	3rd	-0.2%
Principles of Learning and Teaching grades 7–12 (English teachers only)	1st	7.6%
	2nd	8.4%
	3rd	11.6%
	4th	11.6%
	5th	5.4%