The College Board

Research Report No. 2002-12

Examining the Relationship of Content to Gender-Based Performance Differences in Advanced Placement Exams

Gary Buck, Irene Kostin, and Rick Morgan

College Board Research Report No. 2002-12
ETS RR-02-25

Examining the Relationship of Content to Gender-Based Performance Differences in Advanced Placement Exams

Gary Buck, Irene Kostin, and Rick Morgan

College Entrance Examination Board, New York, 2002

Gary Buck is dean of Test Development and Standards Division at the Defense Language Institute.

Irene Kostin is a senior research associate at Educational Testing Service (ETS).

Rick Morgan is a program administrator at Educational Testing Service (ETS).

Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board: Expanding College Opportunity

The College Board is a national nonprofit membership association dedicated to preparing, inspiring, and connecting students to college and opportunity. Founded in 1900, the association is composed of more than 4,200 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three million students and their parents, 22,000 high schools, and 3,500 colleges, through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, visit www.collegeboard.com.

Additional copies of this report (item #996245) 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 help of Sarah Igo, Jennifer Mulligen, Mina Lebitz, and Susan Woodward in developing the coding systems and coding the items. We would also like to acknowledge the help of Lawrence Beaber, Despina Danos, Fekru Debebe, Michael Lapp, Claire Melican, Beth Nichols, and Nancy Olds in assembling the necessary test materials.

Contents

Tables

Abstract

The purpose of this study is to examine the content of the questions in a number of Advanced Placement Examinations and to attempt to identify content that is related to gender-based performance differences. Free-response questions for ten forms of the AP® Exams in U.S. History, European History, Biology, Microeconomics, and Macroeconomics were studied and the multiple-choice items for four forms of AP U.S. History were also studied.

Results indicate that, in the case of U.S. History multiple-choice items, the measures of male and female performance differences are significantly correlated with item content. In a second analysis, the findings for nine of these content categories were replicated in different forms of the exam, suggesting that these results are of generalizable interest. In the case of the free-response questions for each of the five subject areas, the study identified a significant relationship between item content and gender-based performance differences, but there were not enough items for a replication study. Overall, this study suggests that item content is associated with gender-based performance differences.

Introduction

Statement of the Problem

Males and females have different distributions of AP grades for most Advanced Placement Program® (AP) Examinations. The AP Program needs to know whether these performance differences are related to the item content. Knowledge of relationships between item content and performance differences will allow AP to provide exams that are more parallel across forms within and among years.

Background and Literature Review

The literature suggests that males and females tend to have different interests and values. Allport, Vernon, and Lindzey (1970) used an inventory of values to explore male and female differences and found that males scored higher on scales of theoretical, economic, and political values, whereas females scored higher on scales of social, religious, and aesthetic values. Hansen and Campbell (1985) reported gender differences on the Strong Interest Inventory: men scored higher on the scales related to mechanics, the military, athletics, adventure, agriculture, science, sales, mathematics, law/politics, business management, and medical science; whereas women scored higher on the scales related to domestic topics, art, music/drama, writing, social service, office practices, religion, and medical service. More generally, Eccles, Adler, Futterman, Goff, Kaczala, Meece, and Midgley (1983) reviewed numerous studies which showed that females are more likely to be “person oriented” and males are more likely to be “thing oriented” in their interests and values.

Males and females also tend to pursue different interests at school. Regarding SAT® I: Reasoning Test test-takers’ plans for advanced standing in college courses, females more often than males plan for advanced standing in art, English, foreign languages, and the humanities, while males more often plan for advanced standing in the physical sciences and computer science (College Board, 2000). Data from SAT test-takers also show that when asked about their intended college major, relative to males, females more often report intentions to pursue the areas of foreign or classical languages, language and literature, the arts, both visual and performing, and home economics. Males are more likely to report plans to major in mathematics, the physical sciences, engineering, military science, and computer science (College Board, 2000). Astin (1993) noted that in college, females are more likely than males to drop out of law, medicine, and engineering, and are more likely than males to remain in the fields of education, nursing, and psychology. Dwyer and Johnson (1997) reported findings from a study of high school seniors in 1992 that indicated that girls are more likely to win honors or awards in writing, while boys are more likely to win honors and awards in vocational/technical, science/math, and sports areas.

It seems reasonable to expect that these differences would manifest themselves in test performance, and there is considerable evidence to suggest that the content of test items is related to differences between male and female performance in ways that are consistent with gender differences in interests and values. For example, there have been a number of studies that explored the relation of item content to gender-based performance differences on the SAT verbal test (SAT–V). Carlton and Harris (1992) found that on sentence completion and analogy items, females performed differentially better than males on items with content pertaining to human relations and aesthetics/philosophy, and differentially worse on items pertaining to science and practical affairs (which includes sports, economics, business, politics, etc.). For both sentence completion and reading comprehension items, Carlton and Harris report that females performed relatively better than males when people were referred to in the items as compared to when there was no reference to people. Curley and Schmitt (1993) reported that "science, industrial arts, and military terminology, as well as contexts portraying aggression or conflict may negatively affect the performance of females" (p. 2). O'Neill and McPeek (1993) largely concurred: They suggested that on sentence completion and analogy items related to science and practical affairs, males performed better than females, whereas on aesthetics/philosophy and human relationships, females performed better than males. In a similar vein, Ibarra (1997) reported that females performed better than males on verbal questions relating to aesthetics, humanities, and human relationships.

An unpublished study by two of the current authors (Buck and Kostin) explored the relationship of item content to gender-based performance differences on the SAT–V in greater detail than the studies reviewed above. Analogy and sentence completion items from 10 SAT–V forms were used to develop and define a set of content categories. Then analogy and sentence completion items from 11 other forms were used to replicate these findings. The results showed that although the gender-based performance difference of each particular content category was small, taken together the categories accounted for much of the gender-based performance differences of the items. Twelve categories of male-oriented content, and also twelve categories of female-oriented content, were found to meet the replication criteria. These were:

Male-Oriented Content

- hard science
- general science
- applied science
- business and economics
- investigation and problem solving
- competition and conflict
- vehicles
- fame and high achievement
- sports
- political content
- dominance and authoritarian behavior
- physical danger and risk taking

Female-Oriented Content

- human relationships
- feelings and emotions
- personality and behavior
- arts and literature
- domestic items and activities
- personal appearance
- psychology
- verbal aggression
- social service
- health care
- formal education
- religion

Results also indicated that multiple content tends to increase the relationship to gender-based performance differences: e.g., the more the male-oriented content in an item the stronger the performance difference favoring males. Whereas, mixed content had a compensatory relationship: e.g., the relationship of male-oriented content to differential performance was lessened if female-oriented content was also present in an item. The SAT–V items use the multiple-choice (MC) format, and it seems reasonable to hypothesize that similar relationships would be found for the Advanced Placement multiple-choice questions.

The Advanced Placement Examination also uses free-response (FR) questions, and there is evidence to suggest that the specific topic, or content, of such free-response questions might account for at least some of the gender-related differences in performance. For example, Mazzeo, Schmitt, and Bleistein (1993) looked at gender differences in performance on free-response questions of AP Chemistry, Biology, U.S. History, and English Language and Composition examinations. They concluded that the topic of the FR question was very important: "These results . . . suggest that specific topics play a substantial role with respect to the magnitude of sex-related differences" (p. 26). Mazzeo et al. (1993) give examples of free-response questions in the English Language and Composition Exam which were associated with large gender differences: "Questions based on passages related to topics such as patriotism, space satellites, and the ruggedness of the American Prairie produced the largest differences favoring males" (p. 26). It is worth noting that the topics that produced the largest difference favoring males would have been coded as male-oriented in the unpublished study of the

SAT–V referred to above, and other topics that showed a difference favoring females would have been coded as female-oriented. In addition, work by Zwick and Ercikan (1989) on NAEP DIF findings for U.S. History items indicates that males perform differentially better on several items that would have been given male-oriented codes in the SAT–V coding scheme, and females perform better on items that would have received female-oriented codes.

Purpose and Rationale

Males outscore females on almost all AP Exams, except languages, and this difference is especially large in the case of the science exams. This difference is greater on the multiple-choice sections of the test: in 1993 across all subjects, the mean male performance was .20 standard deviations above the mean female performance. On the same examinations, the difference between the mean male and female performance on the free-response questions was .06 standard deviations (Willingham et al., 1997). While differences between male and female performance on both multiple-choice and free-response questions may be due to a variety of factors, the research reviewed in the previous section suggests that the content of the questions may play an important role.

Although the greatest gender differences are found in the MC sections, the impact of the item content of a single item is less important in MC than in free-response (FR). The MC sections use large numbers of items on a variety of topics, and the effect of individual item content is small. However, in the case of free-response essay questions, there are few items, and a single question may constitute a significant proportion of the total examination. Morgan and Maneckshana (1996) looked at data from AP Biology Exams from 1992 to 1995 and found that gender differences on the MC sections were quite similar across all the years. However, when they looked at the FR sections, they found that the 1995 examination had four out of the five FR questions with the largest differences favoring males for that four-year period. Clearly, when examinations have only a small number of questions, the possibility of free-response sections that are not parallel in terms of content coverage is considerable.

The primary purpose of this study is to understand the relationship between item content and gender-based performance differences, in both the MC and the FR sections of the AP Examinations.

Some AP Examinations offer choice among free-response questions. While many feel this has clear beneficial effects from an educational perspective, it leads to a number of problems from a measurement perspective. Obviously, differences in question difficulty can lead to inequities, but also if there are systematic gender-based differences in the questions selected, these could lead to systematic differences in the examination. For both the U.S. History and European History exams the percentage of females and males who select each question will be compared with the content of the questions. We can then explore whether, and to what extent, test-takers choose questions with topics beneficial to their gender.

Method

Sufficient multiple-choice items are available for developing and defining a set of relevant content categories, and also evaluating the generalizability of any findings by determining whether they replicate on a set of different items.

Because the free-response sections use fewer questions, there are not enough questions to both develop a set of content categories and then use these on a different set of items. Therefore, in the case of free-response questions, the purpose is an exploratory identification of relevant categories. Future studies will be needed to address the issue of replication to determine the generalizability of these categories.

To study gender differences in performance on the multiple-choice items, a measure of differential item functioning (DIF), the Mantel-Haenszel D-DIF statistic was used (Mantel and Haenszel, 1959). This method determines DIF by first partitioning the males and females into subgroups with the same total score on the test. It then calculates the ratio of the odds of success of the males over the odds of success of the females, at each score level, and averages these ratios across each score level. The D-DIF index is obtained by multiplying the natural log of the odds ratio estimate by -2.35 to transform the log odds ratio estimate on the ETS delta scale of item difficulty. More difficult items have higher delta values. In this study, a negative Mantel-Haenszel value indicates a DIF favoring males, and a positive value indicates DIF favoring females.

In the case of the free-response questions, there are not enough items to use the Mantel-Haenszel statistic, and so the standardized difference in mean score is used (Morgan and Maneckshana, 1996). The standardized differences were found by taking the difference between the average scores for males and females and dividing the difference by the standard deviation in the population. The greater the absolute value of the number, the greater the difference between the groups. Like the Mantel-Haenszel DIF statistic, a positive standardized difference indicates a difference favoring females, whereas negative values indicate items favoring males.

The standardized difference statistic for the free-response is a measure of performance difference between two groups. Unlike the Mantel-Haenszel statistic, the standardized difference measure does not attempt to match the two groups on a measure of examinee ability. Therefore, while both measures are indicative of difference between two groups, the measures do not have identical meanings.

Multiple-Choice Items

The U.S. History exam was chosen because these multiple-choice items had significant gender-based performance differences. In 1993 the difference between the means for males and the means for females was about one third of a standard deviation (Willingham et al., 1997). Furthermore, the examination covers a wide variety of different topics.

Two forms of the test, 1997 and 1998, with a total of 160 items were used in the exploratory phase of the study. The procedure was as follows:

1) The size of the gender-based DIF was identified using the Mantel-Haenszel statistic.

2) Two researchers (one member of the research team, and an outside content expert) independently developed a set of draft content categories hypothesized to account for performance differences. One of the researchers used the content categories from the SAT–V studies as well as the findings from the literature review on gender differences in values and interests as a guide in developing the categories. Both researchers also used the items to help develop these content categories.

3) The two researchers compared their draft categories, and jointly developed a set of clearly defined content categories, and a system for coding test items on these categories (see the coding scheme in Appendix A).

4) The two researchers separately coded the items from two forms of the test, 1997 and 1998, on these content categories. In applying the coding system, items that contained the particular content were coded as 1, and items that did not contain the particular content were coded as 0. Note that each item could be coded for more than one content category, or for none of the content categories. The two coders then compared their coding, and any errors were identified and corrected. In cases of disagreement, the coding of the content expert was used. This coding was then used in the analysis.

5) The content categories for each item (as noted above, the coding was dichotomous) were correlated with the DIF statistic for each item, and those categories with correlations at a $p < .05$ significance level were included in the replication phase of the study that used a different set of items.

In the replication phase:

1) Items from two other forms of the test, 1996 and 1999, were coded separately by the two coders on the content categories.

2) The two codings were compared in order to identify and correct error, and in cases of disagreement, the coding of the content expert was used.

3) The content categories for each item were correlated with the DIF statistic.

4) Stepwise multiple regression was used, with the DIF statistic as the criterion variable and the content categories as the predictor variables, in order to determine which categories best predict the differential gender-based performance. Stepwise multiple regression was used because of the exploratory nature of the analyses. An additional reason for using multiple regression is that the items can belong to more than one category and other approaches, which require that items be assigned to only a single category, could not be used to estimate the association of content with DIF. There is no theoretical model as to the order of importance of each of the variables. Because of the exploratory nature of the study, the small sample sizes, and multicolinearity of the data, the paper will present the variables that made significant contributions in the stepwise regression and not provide the regression weights.

Free-Response Questions

Consideration of the size of the gender-based differences in AP Examinations, as well as discussions with content experts, suggested that U.S. History, European History, Biology, Microeconomics, and Macroeconomics would be the best subjects to explore gender-based performance differences for the free-response questions. Each examination contains a limited number of FR questions. Questions for 10 years were examined. In the case of U.S. History, Biology, Microeconomics, and

Macroeconomics, these were all questions for the period 1989 to 1998; in the case of European History, the questions for 1989 were not available, and 1988 was used instead, along with all the questions from 1990 to 1998. The overall procedure was as follows:

1) Standardized differences were calculated for all the free-response (FR) questions.

2) A set of content categories was developed.

3) The questions were coded. Due to the length and complexity of both the questions, and the expected response, it was felt that dichotomous coding (1 = present, 0 = not present) would not capture sufficient variation in content for the FR questions. Therefore, we used a short rating scale (3 = the content is dominant, 2 = a moderate amount of content, 1 = a small amount of content, and 0 = the content is not present).

4) The relationship between the item content (as noted above, coded on a scale from 0 to 3) and the standardized difference was examined using correlation and, in most cases, stepwise multiple regression.

5) The mean and standard deviation of the standardized difference for each content category were computed based on the standardized differences of those FR questions that were coded either 3 (dominant content) or 2 (moderate content) for that content category.

There were not enough items to replicate the results by applying the categories to a different set of items. Therefore, the purpose of this part of the study was merely to explore the relationship between content and gender-based performance differences on these FR questions, and to develop a set of categories related to those. It is important to stress that this work should be regarded as hypothesis generation rather than hypothesis testing. It is unknown whether these categories will generalize to other questions.

While the object of these exploratory analyses is to investigate the relationship between item content and gender differences in performance, one secondary research question suggests itself. Is the relationship between item content and gender difference similar for the document-based questions and the standard thematic essay questions? This is discussed below.

Both the U.S. and the European History examinations contain a document-based question (DBQ). The DBQ requires test-takers to evaluate a number of documents (usually between about 8 to 20) which are provided with the question. The AP candidates are expected to use these documents to construct their response to the question. In the case of U.S. History, AP candidates are required to use their own knowledge as well as the content of the documents, whereas in the case of European History, AP candidates do not need any topic-specific knowledge apart from that provided in the documents. For both U.S. History and European History, the standardized differences for the DBQ questions were compared to the standardized differences for the thematic FR questions. For both exams, the difference was statistically significant indicating relatively superior female performance on the DBQ questions as compared to the thematic FR questions. Because of this significant difference, the DBQ questions are analyzed separately for both U.S. History and for European History.

Exam Question Choice

Both the U.S. History and the European History FR sections allow choice. Do males tend to choose questions on which males usually do better, and do females tend to choose items on which females do better?

In order to test that, the difference between the percent of males who responded to a question and the percent of females who responded to the question was computed. A positive value indicated that a greater percentage of males responded to the question, and a negative value indicated that a greater percentage of females responded.

Data Sets

The number of items in the analysis varied from one test to another. Details of each data set are given below.

U.S. History Multiple-Choice Items

In the exploratory phase of the study, 160 U.S. History multiple-choice items were used, taken from the 1997 and 1998 forms of the exam, 80 items from each. One hundred and sixty items from the 1996 and 1999 forms were used in the replication phase.

U.S. History Free-Response Questions

There were 44 thematic U.S. History free-response (FR) questions. One question was dropped because of low response rate, i.e., 3 percent or less for both males and females. From 1989 to 1993, AP candidates were asked to respond to one of five thematic FR questions, and from 1994 to 1998, AP candidates were asked to respond to two thematic FR questions, one question from each of two sets of two questions.

There were 10 U.S. History document-based questions—one question per year, from 1989 to 1998. AP candidates must respond to the DBQ question.

European History Free-Response Questions

There were 58 thematic European History free-response (FR) questions. Two questions were dropped because of low response rate, i.e., 3 percent or less for both males and females. In 1988 and from 1990 to 1993, AP candidates were asked to respond to one of six thematic FR questions, and from 1994 to 1998, AP candidates were asked to respond to two thematic FR questions, one question from each of two sets of three questions.

There were 10 AP European History document-based questions, one for 1988 and one for each year from 1990 to 1998. AP candidates must respond to the DBQ question.

Biology Free-Response Questions

There were 40 Biology FR questions from 1989 to 1998, there being four mandatory questions a year.

Microeconomics Free-Response Questions

There were 26 Microeconomics questions for the years 1989 to 1998. From 1989 to 1992, there were two mandatory questions a year and from 1993 to 1998, there were three mandatory questions a year. Thus, there were 8 questions for the years 1989 to 1992 and 18 questions for the years 1993 to 1998, totaling 26 questions in all.

Macroeconomics Free-Response Questions

There were 18 AP Macroeconomics questions for the years 1993 to 1998 (with three mandatory questions per year). The data set does not include the four questions for the years 1989 to 1992 (there was only one question per year for these years), because a Mann-Whitney U test showed that the standardized differences for these 4 questions differed significantly from the standardized differences for the 18 questions for the years 1993 to 1998, suggesting that these items may be different in some way. (The Mann-Whitney U test is a nonparametric test of difference between independent samples that is primarily used when the samples to be compared are small in size.)

The Coding

Where possible coding was carried out by a content expert and one of the researchers: The function of the latter was largely to identify error. The development of the coding systems varied from one data set to another, and the details are given below.

U.S. History

A graduate student in U.S. history, who had taught AP U.S. History and had participated in scoring AP U.S. History FR questions, was used as the content expert for all U.S. History data sets. As noted earlier, for the multiple-choice coding, a member of the research team used the content categories from the SAT–V studies as well as the findings from the literature review on gender differences in values and interests as a guide in developing the categories for the coding scheme. Both researchers used the items available in the exploratory phase of the study to help develop these content categories.

Ten of the 15 categories in the U.S. History multiple-choice coding scheme are broadly similar to categories in the coding system for the SAT analogies and sentence completion items referred to earlier. For example, U.S. History male-oriented categories that are similar include Politics, Economics, Geography (a subcategory of the SAT category General Science), and War and other Forms of Armed Conflict (a subcategory of the SAT category Competition and Conflict). U.S. History female-oriented categories that are similar include Feelings and Emotions, Arts and Literature, Social Reform Movements (a subcategory of the SAT category, Social Service), and Religion. The coding system for U.S. History multiple-choice items is given in Appendix A.

With only a few minor differences, the categories in the coding system were the same for the U.S. History free-response questions as for U.S. History multiple-choice: Some related codes that were coded separately in the multiple-choice coding scheme were combined for the free-response scheme and some subcategories that were given the same code in the multiple-choice coding scheme were coded separately in the free-response coding scheme. The coding system for U.S. History FR questions is given in Appendix B.

For both history codings (U.S. History and European History), thematic items were double coded, but the coding of the nonexpert was primarily used to identify error. For both history codings, the DBQ's were coded only by the content expert.

European History

A graduate student in European history was used as the content expert. The coding scheme for these FR questions was developed based on the coding system for U.S. History, and also on an examination of the thematic European History FR questions with standardized differences at the two extremes of the distribution. Most of the categories in the European history coding scheme were very similar or identical to the ones in the U.S. History scheme. Two additional male-oriented categories, Science and Other Conflict, were broadly similar to categories used in the scheme for SAT analogy and sentence completion items. Only two categories, the Renaissance and the Enlightenment, were completely new in the European history coding scheme. The coding system for European History FR questions is given in Appendix C.

Biology

The literature review did not offer a sufficient basis to develop hypotheses about which particular biology topics would favor males and which would favor females. Therefore, a different procedure was used. First, 240 multiple-choice items from two forms of the AP Biology Examination were ranked on their DIF values. Then a content expert, a graduate student in Biology, was asked to examine items with extreme DIF values and to hypothesize regarding the categories that differentiated between the items at the two extremes. The content expert then looked at the AP Biology FR questions to see which of these categories were associated with questions that had standardized differences at the two extremes of the distribution and also to see whether the FR questions suggested any additional categories. Based on this procedure, the expert developed the coding scheme. Most of the categories in this coding system were identified in the initial examination of the Biology multiple-choice items and their associated DIF values, that is, independently of the FR questions themselves. The expert then coded all the free-response items using this coding system. None of the researchers had enough knowledge of biology to attempt to apply the coding scheme, and other biology experts were not available. Consequently, a single coder was used. The coding system for Biology FR questions is given in Appendix D.

Economics

The procedure for the two economics exams was similar to that for the biology exam. The multiple-choice items from 3 AP Micro- and 3 AP Macroeconomics examinations totaling 180 items in each of the two sets were ranked separately on their DIF values. A retired teacher of AP Economics, who had participated in scoring AP Economics FR questions, served as the content expert. She was asked to examine items with extreme DIF values in the two sets of multiple-choice items and to hypothesize categories that differentiated between the items at the two extremes. Separate sets of categories were developed for the two exams. The expert then looked at the FR questions to examine which of these categories differentiated between FR questions at the two extremes in terms of their standardized differences and also to see whether the FR questions suggested any additional categories. Based on these procedures, the expert developed two coding systems. As in the case of biology, most of the categories in the coding systems were identified in the initial examination of the multiple-choice items and their associated DIF values. The content expert then used these coding systems to code the AP Economics FR questions. None of the researchers was qualified to code the economics questions, and no other experts were available, so single coding was used. The coding systems are given in Appendices E and F.

Reliability of Coding

In the current study, some questions were double coded and others were single coded. The reliability of these two coding procedures is likely to differ.

It is very difficult to estimate the reliability of a consensus coding system as used in many of the history codes. To do so it is necessary to replicate the whole process with a completely different set of two raters. Typically, an index of coder agreement is used to indicate coding reliability, and in cases where the coding from only one coder is used, this level of agreement is usually reported as an indication of coding reliability. If the coding of two coders is averaged, the reliability is increased, and the index of agreement is adjusted. In the present case of a consensus-based coding system, the discussion between the coders results in the two most important sources of error being reduced and almost eliminated. In the case of random or careless error, coders tend to make different errors, and in a consensus coding scheme, this error is not averaged, but is largely identified and eliminated. The nonrandom error was mostly due to the different coders having slightly different ideas about the meaning of the categories, most especially about the location of the boundaries between the categories. Discussion then becomes a process of clarifying the categories, which greatly increases the reliability of their application, both between and within raters.

However, in the case of those data sets that were single coded, results need to be treated with more caution. While the coders were asked to take care, and were asked to check their work repeatedly, there is no way of knowing how much error is involved. This is not considered a serious problem, as the examination of these FR categories is exploratory and the purpose is hypothesis generation. However, we need to emphasize that only those data that have been double coded can be considered reliable, just as only those results that have been replicated can be considered generalizable.

Table 1

Exploratory Phase: DIF Descriptive Statistics, and Correlations Between Item Content Codings and Mantel-Haenszel DIF Statistics for All 19 Content Categories on 160 U.S. History Multiple-Choice Questions on Forms 1997 and 1998

Content Category	*N-size*	*Mean M-H DIF*	*SD*	*N*	*Correl. with M-H DIF*	*P (1-tailed)*
Male-Oriented Content						
Geography	5	-.712	.344	160	-.242	.001*
Economics	27	.024	.365	160	.038	n/a**
War, Armed Conflict, and Wartime	37	-.325	.436	160	-.325	.000*
Political Parties and Elections	14	-.368	.398	160	-.209	.004
Time Period after WWII	30	-.234	.459	160	-.200	.006
Foreign Relations	16	-.193	.377	160	-.112	.080
Miscellaneous Male-Oriented Content	3	-.877	.689	160	-.230	.002*
Female-Oriented Content						
Feelings and Emotions	4	.745	.330	160	.238	.001*
Arts and Literature	5	.476	.445	160	.173	.015
Marginalized Groups	26	.284	.378	160	.260	.001*
Social Reform Movements	7	.544	.597	160	.234	.002*
The Great Depression of the 1930s	5	.312	.282	160	.116	.073
Religion	4	.520	.455	160	.168	.017
U.S. History 17th and 18th Centuries	34	.078	.442	160	.098	.108
Women	11	.717	.524	160	.389	.000*
Compromise and Cooperation	8	.146	.205	160	.074	.176
Everyday Life	10	.131	.521	160	.076	.171

* $p < .003$

** The correlation is not in the predicted direction, and hence a one-tailed test is not appropriate.

NB. With a Mantel-Haenszel DIF statistic, because the two groups are matched on total score, about half the items yield a negative DIF value, and the remaining items yield a positive value. Thus, the mean of all the DIF values for a given test form will be close to zero.

Results

U.S. History Multiple-Choice Items

There were two phases to the study of the U.S. History multiple-choice items. The first phase was the exploratory study, in which the set of content categories was developed and which used 160 items from the 1997 and 1998 forms of the exam. The second phase was the replication study, in which the generalizability of the content categories was examined by applying them to the items from the 1996 and 1999 forms of the test. These two phases are discussed separately.

The Exploratory Study

Descriptive statistics for the Mantel-Haenszel DIF statistic are given in the first three columns of Table 1 for the items coded on each of the 17 content categories. All but one of the male-oriented categories have negative means, and all the female categories have positive means. This indicates that almost all the categories were related to item performance on these items in the manner predicted. However, the more important question is whether this trend is statistically significant. In order to test this, we correlated the item content codings with the Mantel-Haenszel DIF statistic. The results are given in the last three columns of Table 1. As the categories were developed largely on the basis of work done on the SAT-Verbal, we felt confident in predicting the direction of the correlation, and so a one-tailed test of significance was used. The correlations for 11 categories had significance levels of $p < .05$. In an effort to control for Type I error, the Bonferonni procedure was used to determine the critical probability. Dividing .05 by the number of tests of significance, the critical probability becomes .003. The seven categories with correlations at this latter level of significance are marked with a single asterisk (*).

The Replication Study

The 11 content categories with significance levels of $p < .05$ were included in the replication study. Table 2 gives the descriptive statistics for these on the two test forms for 1996 and 1999. All the male-oriented categories have negative means, and, with the exception of Feelings and Emotions, all the female categories have

Table 2

Replication Phase: DIF Descriptive Statistics, and Correlations Between Item Content Codings and Mantel-Haenszel DIF Statistics for 11 Content Categories on 160 U.S. History Multiple-Choice Questions on Forms 1996 and 1999

		Mean			*Correl. with*	*P*
Content Category	*N-size*	*M-H DIF*	*SD*	*N*	*M-H DIF*	*(1-tailed)*
Male-Oriented Content						
Geography	7	-.311	.585	160	-.125	.058
War, Armed Conflict, and Wartime	38	-.341	.473	160	-.356	.000*
Political Parties and Elections	22	-.234	.507	160	-.175	.013
Time Period after WWII	24	-.340	.594	160	-.400	.000*
Miscellaneous Male-Oriented Content	5	-.602	.647	160	-.203	.005*
Female-Oriented Content						
Feelings and Emotions	2	-.280	.028	160	-.059	n/a**
Arts and Literature	6	.590	.432	160	.218	.003*
Marginalized Groups	29	.165	.496	160	.146	.033
Social Reform Movements	7	.724	.506	160	.291	.000*
Religion	8	.423	.298	160	.182	.011
Women	10	.562	.411	160	.272	.000*

* p < or = .005

** The correlation is not in the predicted direction, and hence a one-tailed test is not appropriate.

positive means. Again with the exception of Feelings and Emotions, the correlations of the content categories with the Mantel-Haenszel DIF statistic are all in the predicted direction. Once more, one-tailed tests of significance are reported. This shows that the correlations with gender-based DIF for 9 out of the 11 content categories had significance levels of p < .05. Only the male-oriented category, Geography, and the female-oriented category, Feelings and Emotions, failed to replicate; all the other categories included in the analysis successfully met the replication criteria. In an effort to control for Type I error, the Bonferonni procedure was used to determine the critical probability. Using this procedure, the critical probability becomes .005. The six categories with correlations at this latter level of significance are marked with a single asterisk (*).

In order to try to identify the most effective categories, we also carried out a backward-stepwise multiple regression. The dependent variable was the Mantel-Haenszel DIF statistic, and the independent variables were the content categories. The analysis was started with all 11 content categories included in the analysis. Variables with a probability of >.05 were deleted one at a time, starting with the variable with the highest probability of occurring by chance. After the variable with the highest probability of occurring by chance was deleted, the remaining variables were entered in a single step and, within this block of variables, the variable with the highest probability of occurring by chance was deleted. The process was repeated, deleting one variable at a time, until all variables had a probability of <.05. In the final regression equation, eight variables were left. Results are given in Table 3.

Table 3

Results of Backward Stepwise Multiple Regression, with only Significant Variables Remaining in the Equation for U.S. History Multiple-Choice Questions

Multiple R	.646
R Square	.418
Adjusted R-Square	.387
Std. Error of Estimate	.419
Content Categories Making Statistically Significant Contributions	
Geography	
War, Armed Conflict, and Wartime	
Time Period after WWII	
Miscellaneous Male-Oriented Content	
Arts and Literature	
Marginalized Groups	
Social Reform Movements	
Women	

These variables constitute the most parsimonious set of predictors of the total gender-based DIF on the multiple-choice U.S. History examination. As a group they are a little different from the list of variables with significant correlations: Feelings and Emotions and Religion did not contribute significant variance. Geography did contribute significant variance to the equation, despite the lack of a significant correlation with the Mantel-Haenszel DIF statistic. The Multiple R was .65, with an R-square of .42, which became .39 when adjusted for sample size. This suggests that these eight variables account for about 40 percent of the variance in gender-based differential performance.

Table 4

Descriptive Statistics, and Correlations with Standardized Difference for All 14 Content Categories on 44 U.S. History Free-Response Questions Administered from 1989 to 1998

		Mean			*Correl. with*	*P*
Content Category	*N-size*	*Sd DIF*	*SD*	*N*	*Sd DIF*	*(1-tailed)*
Male-Oriented Content						
Technology	4	-.173	.083	44	-.274	.036
Economics	14	-.121	.083	44	-.158	.153
War and Military	12	-.114	.081	44	-.186	.113
Wartime Context	2	-.230	.028	44	-.294	.026
Politics	10	-.147	.075	44	-.286	.030
Time Period after WWII	10	-.124	.127	44	-.191	.107
Foreign Relations	10	-.141	.075	44	-.249	.051
All Questions	44	-.081	.112			
Female-Oriented Content						
Feelings and Emotions	3	.007	.047	44	.215	.081
Arts and Literature	1	-.070	n/a	44	.015	.461
Social Concerns	11	-.065	.116	44	.116	.226
Religion	6	.000	.061	44	.319	.017
U.S. History in the 17th and 18th Centuries	10	-.019	.076	44	.208	.033
Women	4	.130	.108	44	.617	.000*
Everyday Life	7	.004	.146	44	.413	.003*

* p < .004

U.S. History Free-Response Questions

In exploring relationships regarding the question content of the U.S. History FR questions, there are two separate issues that were addressed. The first concerns the relationship of question content of the thematic FR questions to gender-based differences in performance; and the second concerns the relationship of the content in the DBQ to gender-based performance differences. These issues will be addressed separately.

Analysis of Question Content

Table 4 provides descriptive statistics for all the 14 content categories for the 44 thematic FR questions (i.e., not the DBQ). The first column gives the number of questions that were coded as having either a 3 or a 2 (dominant content, or moderate content) on that category. The second column gives the mean of the standardized difference for those questions, with the standard deviation of the standardized difference in the next column.

In Table 4, the mean of the standardized differences for all the 44 items is given. This provides a yardstick against which to compare the other means. In the case of U.S. History categories, all seven male-oriented categories have means lower than the overall mean, and all seven female-oriented categories have means higher than the overall mean. Thus all 14 categories have means in the expected direction, suggesting that the categories are related to gender-based differences in performance, in these items.

The last three columns give the correlation between the coding for each category and the standardized difference, for all 44 questions. The final column gives the statistical significance of that correlation. Again, we used a one-tailed test of significance on the grounds that we have strong reason to predict the direction of the correlation, based on our research into the SAT–V and the analysis of the U.S. History multiple-choice items.

The correlations for seven categories had significance levels of p < .05, and one other category, Foreign Relations, came very close to this level. Using the Bonferonni procedure to control for Type I error, the critical probability becomes .004. The two categories with correlations at this latter level of significance are marked with a single asterisk (*).

Table 5

Results of Backward Stepwise Multiple Regression, with only Significant Variables Remaining in the Equation for U.S. History Free-Response Questions

Multiple R	.802
R Square	.644
Adjusted R-Square	.607
Std. Error of Estimate	.070
Content Categories Making Statistically Significant Contributions	
Politics	
U.S. History in the 17th and 18th Centuries	
Women	
Everyday Life	

Table 6

Descriptive Statistics and Nonparametric Correlations of Standardized Difference for All 8 Content Categories with Dominant or Moderate Presence in 10 U.S. History Document-Based Questions Administered from 1989 to 1998

Content Category	*N*	*Mean of St.DIF*	*SD*	*N*	*Rho With St.DIF*	*P (1-tailed)*
Male-Oriented Content						
War and Military	2	-.060	.014	10	-.752	.006*
Politics	1	-.030	-	10	-.126	.365
Time Period after WWII	1	-.010	-	10	-.059	.436
Foreign Relations	2	-.060	.014	10	-.822	.002*
Female-Oriented Content						
Social Concerns	3	.010	.035	10	.258	.236
U.S. History in the 17th and 18th Centuries	1	.000	-	10	.236	.256
Women	1	.160	-	10	.411	.119
Everyday Life	3	.050	.095	10	.463	.089

* p < or = .006

In order to try to identify the most effective categories, we also carried out a backward-stepwise multiple regression. The dependent variable was the standardized difference statistic. The analysis was started with all 14 content categories included in the regression equation. Variables with a probability of >.05 were deleted one at a time, starting with the variable with the highest probability of occurring by chance. After the variable with the highest probability of occurring by chance was deleted, the remaining variables were entered in a single step and, within this remaining block of variables, the variable with the highest probability of occurring by chance was deleted. The process was repeated, deleting one variable at a time, until all variables had a probability of < .05. In the final regression equation, four variables were left. Results are given in Table 5. The Multiple R was .802, with an adjusted R-square of .607, suggesting that these variables accounted for 61 percent of the variance in gender-based performance differences, for these items.

Analysis of the Document-Based Questions

Standardized differences for DBQ questions (mean = .003; S.D. = .064) were found to be significantly larger

Table 7

Descriptive Statistics, and Correlations with the Standardized Difference for All 16 Content Categories on 58 European History Free-Response Questions Administered in 1988 and from 1990 to 1998

Content Category	*N*	*Mean of St.DIF*	*SD*	*N*	*Correl. with St.DIF*	*P (1-tailed)*
Male-Oriented Content						
Science	5	-.062	.093	58	.010	n/a**
20th Century Technology	2	-.180	.085	58	-.144	.140
Economics	17	-.114	.115	58	-.167	.105
War and Military	9	-.166	.078	58	-.335	.005
Wartime Context	2	-.030	.057	58	.045	n/a**
Other Conflict	4	-.153	.084	58	-.157	.120
Politics	21	-.099	.101	58	-.161	.113
Time Period after WWII	7	-.174	.161	58	-.362	.003*
Foreign Relations	13	-.160	.101	58	-.419	.001*
All questions	58	-.076	.119			
Feale-Oriented Content						
Feelings and Emotions	6	-.082	.078	58	-.015	n/a**
Art	7	.016	.097	58	.310	.009
Religion	9	-.024	.082	58	.181	.087
The Renaissance	3	.060	.052	58	.281	.016
The Enlightenment	5	-.054	.081	58	.042	.376
Women	8	.036	.129	58	.435	.000*
Everyday Life	16	-.003	.096	58	.394	.001*

* p < or = .003

** The correlation is not in the predicted direction, and hence a one-tailed test is not appropriate.

(t = 2.286 (df = 52) p = .026, two-tailed) than the standardized differences for the thematic FR questions (mean = -.081; S.D. = .112). Because of this significant difference, the DBQ questions were analyzed separately.

Table 6 gives the descriptive statistics and correlations involving DBQs (with only 10 cases, Spearman's *rho* was used). If no DBQ question was assigned a 2 or a 3 for a category indicating moderate or dominant content, the category was not included in the table. Using the Bonferoni procedure to control for Type I error, the critical probability becomes .006. The two categories with correlations at this latter level of significance are marked with a single asterisk (*).

European History Free-Response Questions

Two separate research questions were addressed with the European History FR questions. The first concerns the relationship of question content of the thematic FR questions to gender-based differences in performance; and the second concerns the relationship of the content in the DBQ with gender-based performance differences. These issues will be addressed separately.

Analysis of the Question Content

Table 7 gives descriptive statistics for all the 16 content categories on the 58 thematic questions (i.e., not the DBQ). The first column gives the number of questions that were coded as having either a 3 or a 2 (dominant content or moderate content) on that category. The second column gives the mean of the standardized difference for those questions, with the standard deviation of the standardized difference in the next column.

In Table 7, the mean of the standardized differences for all the 58 questions is given. This provides a yardstick against which to compare the other means. In the case of these European History categories, seven out of the nine male-oriented categories have means lower than the overall mean, and six out of the seven female-oriented categories have means higher than the overall mean. Thus, 13 out of the 16 categories have means in the expected direction, suggesting that the categories are related to gender-based differences in performance, in these items.

The last three columns give the correlation of the coding for each category with the standardized difference, for all 58 questions. The final column gives the statistical significance of that correlation. Again, one-tailed tests were used. The correlations indicate the effect of each category taken over all the 58 questions. The correlations for seven categories (three male oriented and four female oriented) had significance levels of p < .05. Using the Bonferonni procedure to control for Type I error, the critical probability becomes .003. The four categories with correlations at this latter level of significance are marked with a single asterisk (*).

TABLE 8

Results of Backward Stepwise Multiple Regression, with only Significant Variables Remaining in the Equation for European History Free-Response Questions

Multiple R	.831
R Square	.691
Adjusted R-Square	.640
Std. Error of Estimate	.071
Content Categories Making Statistically Significant Contributions	
20th Century Technology	
Economics	
Politics	
Time Period after WWII	
Foreign Relations	
Art	
Women	
Everyday Life	

A backward-stepwise multiple regression was used to identify the most effective categories. The dependent variable was the standardized difference statistic. The analysis was started with all 16 content categories included in the regression equation. Variables with a probability of >.05 were deleted one at a time, starting with the variable with the highest probability of occurring by chance. After the variable with the highest probability of occurring by chance was deleted, the remaining variables were entered in a single step and, within this remaining block of variables, the variable with the highest probability of occurring by chance was deleted. The process was repeated, deleting one variable at a time, until all variables had a probability of < .05. In the final regression equation, eight variables were left. Results are given in Table 8. The Multiple R is .831 with an adjusted R-square of .640, suggesting that these variables account for 64 percent of the variance in gender-based performance differences, on these items.

Analysis of the Document-Based Questions

Standardized Differences for DBQ questions (mean = .071; SD = .042) were found to be significantly larger (t = 3.867 (df = 66) p = .000, two-tailed) than the standardized differences for the thematic FR questions (mean = -.076; SD = .119). Because of this significant difference, the DBQ questions were analyzed separately. In addition, although the standardized differences for the thematic FR questions for European History and for U.S. History are similar, i.e., mean = -.076, SD = .119 and mean = -.081, SD = .112 respectively, a t-test comparing independent

Table 9

Descriptive Statistics and Nonparametric Correlations of the Standardized Difference for 11 Content Categories with Dominant or Moderate Presence in 10 European History Document-Based Questions Administered in 1988 and from 1990 to 1998

Content Category	*N*	*Mean of St.DIF*	*SD*	*N*	*Rho with St.DIF*	*P (1-tailed)*
Male-Oriented Content						
Science	1	.170	-	10	.824	n/a**
Economics	4	.052	.010	10	-.229	.263
War and Conflict	4	.048	.010	10	-.544	.052
Wartime Context	3	.040	.000	10	-.865	.001*
Other Conflict	5	.050	.010	10	-.403	.124
Politics	5	.046	.009	10	-.679	.015
Foreign Relations	2	.040	.000	10	-.336	.171
Female-Oriented Content						
Feelings and Emotions	8	.062	.025	10	-.008	n/a**
Religion	1	.050	-	10	.273	.223
The Renaissance	1	.090	-	10	.296	.204
The Enlightenment	1	.060	-	10	.454	.094
Women	3	.100	.070	10	.314	.188
Everyday Life	9	.074	.043	10	.635	.024

* p < .004

** The correlation is not in the predicted direction, in which case a one-tailed test is not appropriate.

Table 10

Descriptive Statistics, and Correlations with Standardized Difference for All 12 Content Categories on 40 Biology Free-Response Questions Administered from 1989 to 1998

Content Category	*N*	*Mean of St.DIF*	*SD*	*N*	*Correl with St.DIF*	*P (1-tailed)*
Male-Oriented Content						
Atmospheric Science	1	-.360	n/a	40	-.425	.003*
Basic Chemistry	3	-.200	.142	40	-.169	.148
Applied Biochemistry	3	-.210	.020	40	-.256	.056
Enzymology	2	-.125	.035	40	-.071	.332
Experimental Apparatus	1	-.320	n/a	40	-.352	.013
Structure/Function Relationships	7	-.209	.097	40	-.361	.011
Math	3	-.173	.133	40	-.213	.093
All Questions	40	-.128	.089			
Female-Oriented Content						
Cell Division	1	.040	n/a	40	.308	.027
Experimental Design	5	-.056	.062	40	.293	.033
Genetics and Inheritance	6	-.057	.060	40	.309	.026
Human Physiology	3	-.047	.064	40	.265	.049
Zoology/Classification	2	-.030	.014	40	.257	.055

* p < .004

samples shows that the standardized differences for European History DBQ questions (mean = .071; SD = .042) are significantly larger (t = 2.821, (df = 18), p = .011, two-tailed) than the standardized differences for U.S. History DBQ questions (mean = .003; SD = .064).

Table 9 gives the descriptive statistics and correlations of the 10 DBQ. If no DBQ question was assigned a 2 or a 3 for a category indicating moderate or dominant content, the category was not included in the table. The correlations for three categories had significance levels of p < .05 (with only 10 cases, Spearman's *rho* was used). Using the Bonferonni procedure to control for Type I error, the critical probability becomes .004. The one category with a correlation at this latter level of significance is marked with a single asterisk (*).

The assignment of the male-oriented category, Science to the DBQ question with the largest standardized difference favoring females, i.e., .170, seems to require an explanation. The question asked the AP candidates to "analyze and discuss attitudes and reactions toward the participation of women in the sciences during the seventeenth and eighteenth centuries." Since the DBQ questions for European History only require the AP candidates to use the associated documents in responding to this question, this would minimize the effect of any disadvantage females might have in terms of background knowledge concerning science. The question was also coded for the category, Women, which was one of the strongest female-oriented categories in regard to predicting standardized differences and DIF. Also, the task of analyzing and discussing people's attitudes and reactions provides the basis for assigning the female-oriented category, Feelings and Emotions.

Biology Free-Response Questions

The research question addressed with the Biology FR questions concerns the relationship of question content to gender-based differences in performance.

Analysis of the Question Content

Table 10 gives descriptive statistics for all the 12 content categories on the 40 questions. The first column gives the number of questions that were coded as having either a 3 or a 2 (dominant content or moderate content) on that category. The second column gives the mean of the standardized difference for those questions, with the standard deviation of the standardized difference in the next column.

In Table 10, the mean of the standardized difference for all the items is given. This provides a yardstick against which to compare the other means. Eleven of the 12 content categories have means in the expected direction, which again suggests that overall the categories are related to gender-based differences in performance in these items.

TABLE 11

Results of Backward Stepwise Multiple Regression, with only Significant Variables Remaining in the Equation for Biology Free-Response Questions

Multiple R	.851
R Square	.724
Adjusted R-Square	.653
Std. Error of Estimate	.052
Content Categories Making Statistically Significant Contributions	
Atmospheric Science	
Experimental Apparatus	
Structure/Function Relationships	
Cell Division	
Experimental Design	
Genetics and Inheritance	
Human Physiology	
Zoology/Classification	

The last three columns give the correlation between the coding for each category and the standardized difference, for all 40 questions. The final column gives the statistical significance of that correlation. Again, we used a one-tailed test on the grounds that we had strong reason to predict the direction of the correlation. The correlations for seven categories had significance levels of p < .05. Using the Bonferonni procedure to control for Type I error, the critical probability becomes .004. The one category with a correlation at this latter level of significance is marked with a single asterisk (*).

In order to try to identify the most effective categories, we also carried out a backward-stepwise multiple regression. The dependent variable was the standardized difference statistic. The analysis was started with all 12 content categories included in the analysis. In the final regression equation, eight variables were left. Results are given in Table 11. This shows a Multiple R of .851, with an adjusted R-square of .65, suggesting that the eight variables account for 65 percent of the variance in gender-based performance differences on these Biology free-response items.

Microeconomics Free-Response Questions

The research question addressed with the Microeconomics FR questions concerns the relationship of question content to gender-based differences in performance.

Analysis of the Question Content

Table 12 gives descriptive statistics for all the 9 content categories on the 26 questions. In Table 12, the mean of the standardized difference for all the 26 questions is also given. This provides a yardstick against which to compare the other means. In the case of these Microeconomics categories, eight of the nine content categories have means in the expected direction.

The last three columns show the correlation of the coding for each category with the standardized difference, for all 26 questions. The final column gives the statistical significance of that correlation. Again, we used a one-tailed test on the grounds that we had strong reason to predict the direction of the correlation. The correlations for seven categories had significance levels of p < .05. Using the Bonferonni procedure to control for Type I error, the critical probability becomes .006. The three categories with correlations at this latter level of significance are marked with an asterisk (*). Due to the smaller number of questions in the study, no multiple regression was carried out for Microeconomics.

Table 12

Descriptive Statistics, and Correlations with Standardized Difference for All 9 Content Categories on 26 Microeconomics Free-Response Questions Administered from 1989 to 1998

Content Category	*N*	*Mean of St. DIF*	*SD*	*N*	*Correl. with St. DIF*	*P (1-tailed)*
Male-Oriented Content						
Numerical/Mathematical Manipulation	2	-.275	.078	26	-.449	.011
Graphical Analysis	7	-.177	.057	26	-.444	.011
Cause and Effect Reasoning	5	-.226	.038	26	-.689	.000*
Factor Market	3	-.240	.095	26	-.431	.014
Price Ceilings and Floors	4	-.200	.074	26	-.381	.027
Short- and Long-Run Effects	3	-.250	.020	26	-.532	.003*
All Questions	26	-.129	.083			
Female-Oriented Content						
Recognition and Recall	14	-.076	.047	26	.614	.000*
Application of a Single Concept	8	-.125	.036	26	.147	.237
Externalities	1	-.140	n/a	26	.053	.399

* p < .006

Table 13

Descriptive Statistics, and Correlations with Standardized Difference for All 10 Content Categories on 18 Macroeconomics Free-Response Questions Administered from 1993 to 1998

Content Category	*N*	*Mean of St. DIF*	*SD*	*N*	*Correl. with St.DIF*	*P (1-tailed)*
Male-Oriented Content						
Numerical/Mathematical manipulation	2	-.255	.007	18	-.536	.011
Graphical Analysis	4	-.180	.041	18	-.189	.226
Cause and Effect Reasoning	7	-.160	.043	18	-.107	.337
Money, Banking, and Interest Rates	5	-.216	.050	18	-.658	.002*
Unanticipated Inflation	1	-.240	n/a	18	-.325	.094
All Questions	18	-.159	.062			
Female-Oriented Content						
Recognition and Recall	1	-.030	n/a	18	.261	.147
Application of a Single Concept	4	-.123	.047	18	.312	.104
Effects of Nominal Wages and Labor Productivity	3	-.127	.040	18	.279	.131
Demand, Consumption and the Household Sector	2	-.090	.085	18	.408	.046
Keynesian Theory	6	-.140	.036	18	.244	.165

* p < .005

Macroeconomics Free-Response Questions

The research question addressed with the Macroeconomics FR questions concerns the relationship of question content to gender-based differences in performance.

Analysis of the Question Content

The data set does not include the four questions for the years 1989 to 1992 (there was only one question per year for these years), because a Mann-Whitney U test showed that the standardized differences for these four questions (mean = -.288; SD = .044) differed significantly (U = 1.500, p = .003, two-tailed) from the standardized differences for the 18 questions for the years 1993 to 1998 (mean = -.159; SD = .062).

Table 13 provides descriptive statistics for all the 10 content categories on the 18 questions. The first column gives the number of questions that were coded as having either a 3 or a 2 (dominant content or moderate content) on that category. However, one category, Aggregate Demand, Consumption and the Household Sector, whose highest code was 1 indicating weak content, was also included; this category was included because it had been observed to be strongly associated with DIF favoring females in the Macroeconomics multiple-choice items. The second column gives the mean of the standardized difference for those questions, with the standard deviation in the next column.

In Table 13, the mean of the standardized difference on all the items is given. This provides a yardstick against which to compare the other means. In the case of these Macroeconomics categories, all 10 content categories have means in the expected direction.

The last three columns show the correlation between the coding for each category and the standardized difference for all 18 questions. The final column gives the statistical significance of that correlation. Again, we used a one-tailed test on the grounds that we had strong reason to predict the direction of the correlation. The correlations for three categories had significance levels of $p < .05$. Using the Bonferonni procedure to control for Type I error, the critical probability becomes .005. The one category with a correlation at this latter level of significance is marked with a single asterisk (*). Due to the smaller number of questions in the study, no multiple regression was carried out for Macroeconomics

Exam Question Choice

AP candidates are allowed choice on two of the AP Examinations in the study, U.S. History and European History. To examine whether AP candidates tend to choose topics on which their gender tends to do better, the variable, Chosen Preference, was correlated with the standardized difference. The variable Chosen Preference has a positive value if males tend to choose a question more than females, and a negative value if females choose a question more. The standardized differences will be positive if females outscore males and negative if males outscore females. Therefore, if AP candidates tend to choose topics on which their gender has a performance advantage, then the correlation between Chosen Preference and standardized difference would be negative.

Examining the correlation between Chosen Preference and standardized difference for U.S. History we find a correlation (Spearman's *rho*) of -.360 ($p = .017$) and for European History we find a correlation (Spearman's *rho*) of -.363 ($p = .005$), indicating that there is a significant correlation between test-takers preferences and questions on which they are likely to do better.

Summary and Discussion

Multiple-Choice Items

This study shows that there is a relationship between size of DIF value for AP U.S. History multiple-choice items and their topics and that the relationship appears to be relatively stable across different forms of the test.

The content categories that had significant differences in the replication study as well as the exploratory study can be regarded as having generalizable effects that are likely to occur in other multiple-choice items in other forms of the exam, and probably in other subject areas. Results show that males tend to perform better on items relating to war, armed conflict, wartime, political parties, elections, the time period after World War II, and a number of other topics that males have been found to be more interested in or to value more. Females tend to perform better on items relating to arts and literature, marginalized groups, social reform movements, religion, and women.

Although these categories do not account for all the gender-based performance differences on the multiple-choice items in the U.S. History examination, the multiple regression shows they do account for approximately 40 percent of the variance. The findings can provide test developers with information that should aid in creating exam forms that are more parallel in regard to gender-based differential performance.

Free-Response Questions

The analysis also suggests that the content of the thematic free-response questions is related to gender-based performance differences. The study looked at two different history examinations, U.S. History and European History, one science examination, Biology, and two social sciences, Microeconomics and Macroeconomics.

In the case of the U.S. History FR questions, 14 categories were explored, and in all cases questions that were coded for each content category had means in the expected direction. That is, all questions with male-oriented content had means below the overall mean, and all questions with female-oriented content had means above the overall mean. Of these, four categories, Politics, U.S. History in the Seventeenth and Eighteenth Centuries, Women, and Everyday Life accounted for 61 percent of the variance in our measure of gender-based differences, the standardized difference. Analysis of the European History questions produced a similar result. Thirteen of the 16 content categories explored had means in the expected direction, and in the multiple regression, eight categories, Twentieth Century Technology, Economics, Politics, Time Period after WWII, Foreign Relations, Art, Women, and Everyday Life accounted for 64 percent of the variance in standardized difference.

The literature indicates that females tend to have different values and interests than do males, and this could explain the significant relationships for categories such as Politics, Twentieth Century Technology, Economics, Art, and Everyday Life.

The study also identified relationships between content and gender based performance differences for Biology. Twelve categories were identified as likely to be related to gender-based performance differences and 11 of these had means in the expected direction. A multiple regression showed that eight categories, Atmospheric Science, Experimental Apparatus, Structure/Functional Relationships, Cell Division, Experimental Design, Genetics and Inheritance, Human Physiology, and Zoology Classification accounted for 65 percent of the variance in the standardized difference.

Some explanation of the results for Biology may be found in the work of Eccles et al. (1983), who reviewed a number of studies which found that females are more likely to be "person oriented" and males are more likely to be "thing oriented" in their interests and values. This might help to account for some of the findings in Biology: categories that are more closely related to people, such as Human Physiology and Genetics and Inheritance along with the associated category of Cell Division, have correlations favoring females, while categories that have less to do with people, such as Atmospheric Science and Experimental Apparatus, have correlations favoring males.

Willingham, Cole et al., (1997) looked at performance on a variety of tests taken by twelfth-grade students and found that the largest differences favoring females were in tests of writing. This finding might help account for the correlation, favoring females, of standardized difference with Experimental Design, a code that was assigned to Biology FR questions that ask the AP candidate to design an experiment. Experimental design requires the organization of materials according to a specified format, a task which might be facilitated by superior writing skills.

The results for Biology might help explain why the standardized differences for the four 1995 Biology FR questions were more negative than the standardized differences for all but one out of a total of 20 Biology FR questions examined by Morgan and Maneckshana (1996). Each of the four 1995 questions was coded for one male-oriented code while none of these four questions was coded for any female-oriented code.

Analysis of the two economics examinations showed similar results. In the case of Microeconomics, nine categories were identified as potentially important, and eight of these had means in the expected direction. In the case of Macroeconomics, 10 categories were identified and examined, and all 10 had means in the expected direction.

The correlations of standardized difference with Numerical and Mathematical Manipulation, favoring males, for both Micro- and Macroeconomics are in line with studies by researchers such as Hyde, Fennema, and Lamon (1990) who have reported findings showing superior performance by males in mathematics.

The obvious conclusion from these analyses is that the content of the thematic FR questions does have a relationship with gender-based performance differences. As we have stressed, these results need to be replicated on other items on other forms of the test before we can make strong claims about their generalizability. However, the categories did not come simply from examination of the items themselves, but from the research literature, from other studies carried out by the researchers, and from examination of items other than the FR items; furthermore, the conclusions are consistent with findings in the research literature. This provides strong evidence to suggest that many of these categories will prove to have generalizable effects.

The Document-Based Questions

In the case of the DBQ, it is more difficult to reach firm conclusions regarding the effect of content. The number of questions is so small, making it difficult to identify less powerful effects. More research is clearly needed into how these questions work before any strong conclusion can be drawn.

One can conclude, however, that for both U.S. History and for European History, gender differences in performance on the DBQ differ significantly from gender differences in performance on the thematic FR questions. Relative to males, females perform better on DBQ questions than on thematic FR questions. Since performance on DBQ questions depends more on writing skill as compared to the thematic FR questions, this relatively superior performance of females on DBQ questions may be due to females' generally superior writing skills.

The writing skills of females might also account for their relatively superior performance on European History DBQ questions, compared to U.S. History DBQ questions. Writing skill may be more important for the European History DBQ since the AP candidates are only required to use information provided in the documents, whereas the U.S. History DBQ requires the AP candidates to respond using their topic-specific knowledge as well as the information in the documents.

Exam Question Choice

Results indicate that males and females do tend to choose questions with content on which their gender tends to perform better. This has important implications for test development. If there are known gender-related differences in performance on free-response items that deal with certain topics, and if an examination has only a small number of free-response questions, one obvious strategy test developers might consider would be to include more questions in the examination and allow test-takers to choose topics that interest them. Of course, it is important to ensure that the questions are of equal difficulty, and there is a considerable danger that some test-takers would make bad choices, but nevertheless, the results suggest that this is worth further exploration as a means of making the AP Examination fairer.

Implications for Advanced Placement Program

Clearly the first consideration in determining the appropriacy of AP Examination content is the purpose of the examination, and the nature of the construct that is measured. Decisions on these issues need to be made by the AP Program, and perhaps the wider educational community, and the authors do not feel qualified to comment on whether or not particular content is appropriate for a particular examination. If a content domain is important for construct validity, and if one group of AP candidates performs significantly better on questions from that content domain, we do not feel a need to change the content. We are not suggesting that all groups need to have similar performance on all examinations. However, this research suggests strongly that the content is significantly related to performance differences between subgroups, and the implication is that the AP Program should consider very carefully what is the appropriate content domain for each examination.

However, even when the content domain is well defined, test developers have a wide choice of topics on which to base their items as they sample from that domain. It is very important that they sample these topics with care. In the case of the MC items, it is important that they broadly cover the content domain in a fair and representative manner. In the case of FR examinations, care must be taken to ensure that the small number of questions does not disadvantage one group. The results indicate that providing choice may mitigate some of the performance differences between males and females. It is suggested that test developers balance their FR sections using questions from both male oriented and female oriented content areas. This would reduce the risk of having exam forms that are far from parallel in relation to gender-based score differences.

This research is to a considerable extent exploratory. The issues are complex, and this one study only begins to address the complexities. We recommend strongly that the AP Program carry out more research into the relationship between question content and differences in subgroup performance. First, we recommend studies to replicate these results related to gender-based differences, and second, we recommend research to explore the relationships between content and performance differences for other subgroups.

References

Allport, G., Vernon, P. & Lindzey, G. (1970). *Study of Values—A scale for measuring the dominant interests in personality* (3rd ed.). New York: Houghton Mifflin.

Astin, A. W. (1993). *What matters in college? Four critical years revisited*. San Francisco: Jossey-Bass.

Buck, G. & Kostin, I. (1998). *Exploring the effects of item content on the gender-related impact of SAT–Verbal Sentence Completion and Analogy item*. Unpublished Report.

College Board. (2000). College Bound Seniors 2000. New York: College Entrance Examination Board.

Carlton, S. T. & Harris, A. M. (1992). *Characteristics associated with differential item functioning on the scholastic aptitude test: Gender and majority/minority group comparisons*. (ETS Research Report RR-92-64). Princeton, NJ: Educational Testing Service.

Curley, W. E. & Schmitt, A. P. (1993). *Revising SAT–Verbal items to eliminate differential item functioning*. (ETS Research Report RR-93-61). Princeton, NJ: Educational Testing Service.

Dwyer, C. A. & Johnson, L. M. (1997). Grades, accomplishments, and correlates. In Willingham, W. & Cole, N. (eds.). *Gender and fair assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Eccles, J.S., Adler, T.F., Futterman, R., Goff, S.B., Kaczala, C.M., Meece J.L. & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 75–146). San Francisco: Freeman.

Hansen, J. & Campbell, D. (1985) Manual for the SVIB-SCII. Strong-Campbell interest inventory—Form T325 of the Strong Vocational Interest Bland (4th ed.) Stanford, CA: Stanford University.

Hyde, J. S., Fennema, E. & Lamon, J. S. (1990). Gender differences in mathematical performance: A meta-analysis. *Psychological Bulletin*, 105, 198–214.

Ibarra, R. (1997, May). Cultural context: A new cognitive model for examining ethnic and gender-related DIF. Paper presented at Educational Testing Service, Princeton.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute,* 22, 719–748.

Mazzeo, J., Schmitt, A. & Bleistein, C. (1993). *Sex-related performance differences on constructed-response and multiple-choice sections of Advanced Placement examinations*. (ETS Research Report RR-93-5). Princeton, NJ: Educational Testing Service.

Morgan, R. & Maneckshana, B. (1996, April). The psychometric perspective: Meeting four decades of challenge. In *Lessons learned from 40 years of constructed response testing in the Advanced Placement Program*. Symposium conducted at the NCME Conference.

O'Neill, K. A. & McPeek, W.M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–280). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Willingham, W., Cole, N., Lewis, C. & Leung, S. (1997). Test Performance. In Willingham, W. & Cole, N. (eds.) *Gender and fair assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Zwick, R. & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP History Assessment. *Journal of Educational Measurement*, 26, 55–66.

Appendices: Definition of Content Categories and Coding Instructions

Appendix A: Content Categories for U.S. History Multiple-Choice Items

Listed below are content characteristics as we have defined and operationalized them. Every item should be carefully evaluated for each of these codes. If a specific content code is determined to be present, a 1 [one] is assigned. If a specific content code is determined not to be present, a 0 [zero] is assigned. More than one code will often be assigned for each item; on the other hand, some items will have no code assigned to them. Unless otherwise noted, the codes apply only to the content in the stem and the key.

The coding for "EXCEPT" items is determined primarily by the content in the item stem. However, the item may be coded for another topic if at least two of the four distracters are related to another code.

After each code's description, an example is given. All examples are taken from the disclosed 1996 AP U.S. History form.

Content Hypothesized to Differentially Favor Males

Geography. The key for items coded for geography refers to some geographical location. At least three of the four distracters should also refer to different geographical locations. This code is NOT applied to references to broad geographical regions, e.g., European.

Example:

(Referring to a map on the page) The area marked X on the map was part of

(A) Massachusetts' Western Reserve

(B) the Northwest Territory

(C) the Louisiana Purchase*

(D) the Mexican Cession

(E) the Oregon Country

Economics. Item content coded for economics includes general references to the economy or to economics as well as references to the more specific topics of business and the banking system. This code does not include item content that focuses on workers or organizations that represent workers such as labor unions.

Example:

Settlers who established the British colony in Virginia during the seventeenth century were primarily seeking to

(A) recreate an Old World feudalistic society in the New World

(B) create a perfect religious commonwealth as an example to the rest of the world

(C) create a refuge for political dissidents

(D) profit economically*

(E) increase the glory of Great Britain

War, other forms of armed conflict, and wartime. This code is assigned to item content that refers not only to war and content directly related to war, e.g., troops, weapons, public opinion about war, etc. but also to nonmilitary events that occur during wartime. This code is also assigned to events directly leading up to and precipitating war and to events thought of as the primary triggers of war as well as to events that directly follow or are a consequence of war.

Item content referring to threats of military action either by the United States or by potential foes is also coded. Also item content referring to forceful expansion is coded. Smaller military skirmishes are also coded. (In such cases, both parties in the encounter are armed.)

Example:

In 1861 the North went to war with the South primarily to

(A) liberate the slaves

(B) prevent European powers from meddling in American affairs

(C) preserve the Union*

(D) avenge political defeats and insults inflicted by the South

(E) forestall a Southern invasion of the North

Politics. Item content coded for politics refers to any content concerning electoral politics (in addition to any content concerning electoral politics, this code is assigned if the question uses the term "politics" or "political").

Example:

Jimmy Carter and Ronald Reagan were similar as presidential candidates in that both

(A) articulated the public's desire for less involvement in foreign affairs

(B) capitalized on their status as Washington outsiders*

(C) promised Congress increased control over domestic matters

(D) renounced private fund-raising in support of their campaigns

(E) had built national reputations as legislators

Time Period after WWII. Item content dealing with the situation or events in the period after World War II is assigned this code.

Example:

In negotiations to end the Cuban Missile Crisis, President Kennedy promised to

(A) send economic aid to Cuba under the Alliance for Progress

(B) allow Cuban Propaganda in Latin America

(C) reduce the number of United States missiles on the North American continent

(D) refrain from a military invasion of Cuba*

(E) establish a quota system for Cuban refugees to the United States

Foreign Relations. Item content dealing with the interactions between the United States and foreign countries aside from war or interactions leading up to war is assigned this code. This code includes relations between Britain and the colonists.

Example:

Jefferson's purchase of Louisiana had its origins in his desire to

(A) remove the French from forts along the Mississippi valley

(B) acquire a port to provide an outlet for western crops*

(C) acquire territory for the expansion of slavery

(D) oppose New England Federalism

(E) demonstrate friendship for the French in the Napoleonic Wars

Miscellaneous Male-Oriented Content. This code covers male-oriented content that is only represented by single items in the multiple-choice exam. This content includes technology, machines, inventions, transportation, and sports.

Example:

The assembly-line production of Henry Ford's Model T automobile resulted in which of the following by the end of the 1920s?

(A) A sharp decrease in railroad passenger traffic

(B) The federal government's abandonment of research on air travel

(C) The development of a large international market for American automobiles

(D) Widespread purchase of automobiles by average American families*

(E) Construction of the federal interstate highway system

Content Hypothesized to Differentially Favor Females

Feelings and Emotions. Items are coded for feelings and emotions if the *key* refers to some feeling or emotion and/or if at least three of the four distracters refer to some feeling or emotion.

Example:

Which of the following led immediately and directly to Theodore Roosevelt's issuance of the Roosevelt Corollary to the Monroe Doctrine?

(A) Pancho Villa's armed raids into Texas and New Mexico

(B) General Augusto Sandino's insurrection against American troops occupying Nicaragua

(C) The arrest of an unarmed party of American sailors in Tampico, Mexico

(D) American concern that a Japanese syndicate would attempt to purchase land near the Panama Canal

(E) American fear that financial instability in the Dominican Republic would lead to European intervention*

Arts and Literature. Items coded for arts and literature include content referring to either literary works or to the arts such as music or painting; also items referring to speeches or orations that have some literary merit are assigned this code.

Example:

Which of the following best characterizes the writers associated with the literary flowering of the 1920s, such as Sinclair Lewis and F. Scott Fitzgerald?

(A) Sympathy for Protestant fundamentalism

(B) Nostalgia for the "good old days"

(C) Commitment to the cause of racial equality

(D) Advocacy of cultural isolationism

(E) Criticism of middle-class conformity and materialism*

Social Concerns. In general, the following subcodes refer to movements and groups (other than political parties) that either aim to improve conditions in society as a whole or aim to improve the situation of the group members themselves. These codes also, generally, include references to the more disadvantaged or needy members of U.S. society.

Marginalized groups and movements either for or against these groups. This code is assigned to items that deal with the topic of slavery, efforts to abolish slavery other than war, and later efforts to improve the situation of Black Americans in society, i.e., the civil rights movement. This code is NOT assigned to items that refer to more positive historical events concerning Black Americans such as an increase in the free Black population in the eighteenth century or to a major Black American sports figure.

This code is also assigned to items that deal with the topic of immigrants, immigration, and movements relating to these groups.

Example:

All of the following account for nativist sentiment against the "new immigrants" of the late nineteenth century EXCEPT that the immigrants

(A) practiced different religions

(B) had different languages and cultures

(C) were willing to work for lower wages than were native-born workers

(D) were not familiar with the U. S. political system

(E) dominated the professions of law, medicine, and engineering*

Social Reform Movements. This code is assigned to social reform movements that do NOT pertain to marginalized groups. This code is NOT assigned to movements concerning the environment or conservation.

Example:

The goals of educational reformers in the antebellum years included all of the following EXCEPT

(A) Compulsory school-attendance laws

(B) The use of state and local tax money to finance public education

(C) The establishment of teacher-training schools

(D) A standardized length for the school year

(E) Federal financing of secondary education*

The Great Depression of the 1930s. This topic comes under "Social Concerns" because it was a time when large numbers of people in the U.S. were in need of help.

Example:

President Herbert Hoover approached the task of caring for unemployed workers during the Great Depression by

(A) Emphasizing the importance of private charities*

(B) Asking large corporations to hire war veterans

(C) Relying on the services of federal welfare agencies

(D) Enlarging the federal government's payroll

(E) Reactivating the dole

Religion. Items with any content in the stem or key that refers to religion are assigned this code. This code is also NOT assigned to items dealing with religious rhetoric.

Example:

Which of the following was true of the first Great Awakening?

(A) It primarily affected church congregations in towns and cities.

(B) Cotton Mather was one of its most famous preachers.

(C) It was denounced by Jonathan Edwards.

(D) It was primarily a southern phenomenon.

(E) It resulted in divisions within both the Congregational and the Presbyterian churches.*

U.S. History in the Seventeenth and Eighteenth Centuries. Item content dealing with the situation or events in the United States during the seventeenth and eighteenth centuries.

Example:

The Proclamation of 1763 did which of the following?

(A) Introduced a tax on tea

(B) Prohibited colonists from producing iron for the American market

(C) Forbade all colonial trade with the French West Indies

(D) Set a boundary along the crest of the Appalachians beyond which the English colonists were forbidden to settle*

(E) Announced the reorganization of the colonial office under Parliament, rather than directly under the King-in-Council

Women. Content in which the stem and/or the key involve women is assigned this code.

Example:

In the early 1830s, the majority of workers in the textile mills of Massachusetts were

(A) young unmarried women from rural New England.*

(B) newly arrived immigrants from Ireland.

(C) men who were heads of households.

(D) married women whose children were of school age.

(E) free African Americans from urban areas.

Compromise cooperation and attempts to avoid conflict. This code is assigned to item content referring to various kinds of positive relations among individuals and nations, attempts to reach agreement, formal "compromises," and attempts to avoid or lessen the chances of conflict, especially of armed conflict.

Example:

During the 1930s, the Roosevelt administration did which of the following?

(A) Ceded the Panama Canal Zone to Panama.

(B) Granted immediate independence to the Philippines.

(C) Formally renounced the right to intervene in Latin America.*

(D) Established the Organization of American States.

(E) Held a referendum in Puerto Rico on the commonwealth's entry to the Union.

Everyday Life. This code is assigned to content about settlement, household, working conditions, education, and other topics relating to daily life (excluding slavery) which occur primarily during peacetime.

Example:

Which of the following statements about American cities between 1890 and 1930 is correct?

(A) Area of residence increasingly became an indicator of social class.*

(B) Poor people moved to the outskirts of cities.

(C) Industries shifted from the cities to the suburbs.

(D) Widespread racial integration of communities occurred.

(E) Neighborhoods lost their ethnic identification.

Appendix B: Content Categories for U.S. History Free-Response Questions

Listed below are content characteristics as we have defined and operationalized them. There will be two sets of codes for the essay questions. First, codes will be assigned to the essay questions themselves; then, codes will be assigned separately to the scoring guides that are associated with the essay questions.

Each code will be evaluated according to the following rating scale: 3 = dominant content; 2 = moderate content; 1 = weak content; 0 = not present. The essay question itself will be assigned a 3 if the question requires that all or nearly all of the essay should deal with this topic; in the case of the associated scoring guide, if about 70 percent to 100 percent of the topics listed deal with this topic. The essay question will be assigned a 2 if the code refers to one of a list of two or three topics that the question requires the AP candidate to write about; in the case of the associated scoring guide, if about 30 percent to 69 percent of the topics listed deal with this topic. The essay question will be assigned a 1 if the code refers to one of a list of more than three topics the question requires the AP candidate to write about; in the case of the scoring guide, up to about 29 percent of the topics listed deal with this topic.

Content Hypothesized to Be Negatively Associated with Standardized Differences (favoring males)

Technology. Essay questions are assigned this code if they refer to technology or to some similar concept (for example, inventions and improvements in transportation).

Example:

The reorganization and consolidation of business structures was more responsible for late nineteenth-century American industrialization than was the development of new technologies.

Assess the validity of this statement with specific reference to business structures and technology between 1865 and 1900.

Economics. Essay questions coded for economics include general references to the economy or to economics as well as references to the more specific topics of business and the banking system. This code does *not* include item content that focuses on workers or organizations that represent workers such as labor unions.

Example:

Developments in transportation rather than in manufacturing and agriculture sparked American economic growth in the first half of the nineteenth century.

Assess the validity of this statement.

(This question is also coded 3 for 04 Economics and 3 for 03 Technology.)

War and Military. This code is assigned to essay questions that refer to war, forceful expansionism, or to military aggression in general. This code is also assigned to essay questions concerning factors leading up to and precipitating war as well as to developments that directly follow or are a consequence of war.

Example:

Assess the relative influence of THREE of the following in the American decision to declare war on Germany in 1917.

German naval policy

American economic interests

Woodrow Wilson's idealism

Allied propaganda

America's claim to world power

Wartime Context. This code is assigned to questions that refer to nonmilitary developments that occur during wartime, such as economic or social change.

Example:

"Foreign affairs rather than domestic issues shaped presidential politics in the election year 1968."

Assess the validity of this statement with specific reference to foreign and domestic issues.

Politics. This code is assigned to essay questions that concern electoral or party politics or, in most cases, to anything specifically referred to as political. This code is NOT assigned to questions where the term "political" refers to constitutional issues.

Example:

Evaluate the relative importance of domestic and foreign affairs in shaping American politics in the 1790s.

Time Period after WWII. Essay questions dealing with the situation or events in the period after World War II are assigned this code.

Example:

"Vice Presidents who have succeeded to the presidency on the death of the President have been less effective in their conduct of domestic AND foreign policy than the men they replaced."

Assess the validity of this statement for any TWO of the following pairs.

William McKinley and Theodore Roosevelt

Franklin D. Roosevelt and Harry S. Truman

John F. Kennedy and Lyndon B. Johnson

Foreign Relations. Essay questions dealing with the interactions between the United States and foreign countries aside from war are assigned this code. These questions use terms such as "foreign issues," "foreign policy," "foreign relations," or "diplomacy." In addition are questions that deal with the related topics of the United States as a world power, threats to its position of world power, e.g., communism, and reactions within the United States to such threats, e.g., McCarthyism.

Example:

In 1945 Winston Churchill said that the United States stood at the summit of the world. Discuss the developments in the 30 years following Churchill's speech, which called the global preeminence of the United States into question.

Content Hypothesized to Be Positively Associated with Standardized Differences (favoring females)

Feelings and Emotions. Questions are assigned this code if they include a term or concept that refers, at least in part, to feelings or emotions, e.g., attitudes.

Example:

The Bill of Rights did not come from a desire to protect the liberties won in the American Revolution, but rather from a *fear* of the powers of the new federal government.

Assess the validity of the statement.

Arts and Literature. Essay questions coded for arts and literature include content referring to either literary works or to the arts such as music or painting.

Example:

"Although American writers of the 1920s and the 1930s criticized American society, the nature of their criticisms differed markedly in the two decades."

Assess the validity of this statement with specific reference to writers in both decades.

Social Concerns. This code is assigned to content about social movements, the groups that these movements are designed to help, and social reform including the following subtopics: (1) Marginalized groups and movements either for or against these groups; this included references to slavery, efforts to abolish slavery other than war, and later efforts to improve the situation of Black

Americans in society, i.e., the civil rights movement; references to immigrants, immigration, and movements relating to immigrants; and references to Native Americans; (2) Reform movements whose aim is to improve society in a number of areas; and (3) The Great Depression of the 1930s—this topic comes under "Social Concerns" because it was a time when large numbers of people in the United States were in need of help.

Example:

In what ways did the early nineteenth-century reform movement for abolition and women's rights illustrate both the strengths and the weaknesses of democracy in the early American republic?

Religion. Essay questions that refer to religion are assigned this code.

Example:

Analyze the extent to which religious freedom existed in the British North American colonies prior to 1700.

U.S. History in the Seventeenth and Eighteenth Centuries. Item content dealing with the situation or events in the U.S. during the seventeenth and eighteenth centuries.

Example:

For the period before 1750, analyze the ways in which Britain's policy of salutary neglect influenced the development of American society as illustrated in the following.

Legislative assemblies

Commerce

Religion

Women. Essay questions whose content involves women are assigned this code.

Example:

From the 1840s through the 1890s, women's activities in the intellectual, social, economic, and political spheres effectively challenged traditional attitudes about women's place in society.

Assess the validity of this statement.

Everyday Life. Questions are assigned this code if they refer to household or other labor, settlement life, social patterns, education or daily life or, in general, if the word "social" appears in the question. This code is NOT assigned if the word "social" refers primarily to social reform movements; in the latter case code "Social Concerns" would apply.

Example:

Analyze the ways in which the Great Depression altered the American social fabric in the 1930s.

Appendix C: Content Categories for European History Free-Response Questions

Listed below are content characteristics as we have defined and operationalized them. There will be two sets of codes for the essay questions. First, codes will be assigned to the essay questions themselves; then, codes will be assigned separately to the scoring guides that are associated with the essay questions.

Each code will be evaluated according to the following rating scale: 3 = dominant content; 2 = moderate content; 1 = weak content; 0 = not present. The essay question itself will be assigned a 3 if the question requires that all or nearly all of the essay should deal with this topic; in the case of the associated scoring guide, if about 70 percent to 100 percent of the topics listed deal with this topic. The essay question will be assigned a 2 if the code refers to one of a list of two or three topics that the question requires the AP candidate to write about; in the case of the associated scoring guide, if about 30 percent to 69 percent of the topics listed deal with this topic. The essay question will be assigned a 1 if the code refers to one of a list of more than three topics the question requires the AP candidate to write about; in the case of the scoring guide, up to about 29 percent of the topics listed deal with this topic.

Content Hypothesized to Be Negatively Associated with Standardized Differences (favoring males)

Science. Essay questions are assigned this code if they refer to science.

Example:

Describe the new astronomy of the sixteenth and seventeenth centuries and analyze the ways in which it changed scientific thought and method.

Twentieth Century Technology. Essay questions are assigned this code if they refer to technology of the twentieth century or to some similar concept (for example, inventions and improvements in transportation).

Example:

Identify four specific changes in science and technology and explain their effects on Western European family and private life between 1918 and 1970.

Economics. Essay questions coded for economics include general references to the economy or to economics (including socioeconomics) as well as references to the more specific topics of business and the banking system. This code does *not* include item content that focuses on workers or organizations that represent workers such as labor unions.

Example:

Analyze the changes in the European economy from about 1450 to 1700 brought about by the voyages of exploration and by colonization. Give specific examples.

War and Military. This code is assigned to essay questions that refer to war, forceful expansionism, or to armed conflict in general.

Example:

Account for the responses of the European democracies to the military aggression by Italy and Germany during the 1930s.

Wartime Context. This code is assigned to questions that refer to nonmilitary developments, such as economic or social change, that occur during wartime or during periods characterized by violence or armed conflict.

Example:

Identify the major social groups in France on the eve of the 1789 Revolution. Assess the extent to which their aspirations were achieved in the period from the meeting of the Estates-General (May 1789) to the declaration of the republic (September 1792).

Other Conflict. This code is assigned to questions that refer to conflicts between groups *other than nation states* that may or may not have resulted in armed conflict.

Example:

Describe and analyze the ways in which sixteenth-century Roman Catholics defended their faith against the Protestant Reformation.

Politics. This code is assigned to essay questions that use the terms “politics” or “political.” This includes political practice and intellectual political history (terms such as “democracy” and “monarch”; specific political aspects of concepts such as “nationalism,” “liberalism,” “Marxism,” etc.)

Example:

Analyze the military, political, and social factors that account for the rise of Prussia between 1640 and 1786.

Time Period after WWII. Essay questions dealing with the situation or events in the period after World War II are assigned this code.

Example:

Describe and analyze the resistance to Soviet authority in the Eastern bloc from the end of the Second World War through 1989. Be sure to include examples from at least two Soviet satellite countries.

Foreign Relations. Essay questions dealing with the relationships or interactions between countries (or between countries and continents) aside from war or other military encounters are assigned this code. These questions frequently use terms such as “foreign issues,” “foreign policy,” “foreign relations,” or “diplomacy.” This code is also assigned to questions concerning European colonialism.

Example:

Compare and contrast the relationships between the great powers and Poland in the periods 1772–1815 and 1918–1939.

Content Hypothesized to Be Positively Associated with Standardized Differences (favoring females)

Feelings and Emotions. Questions are assigned this code if they include a term or concept that refers, at least in part, to feelings or emotions, e.g., attitudes.

Example:

Compare and contrast the attitudes toward science and technology held by Enlightenment thinkers with the various attitudes held by European artists and intellectuals in the twentieth century.

Art. Questions that inquire about the visual arts (including architectural design) are assigned this code.

Example:

Discuss how Renaissance ideas are expressed in the Italian art of the period, referring to specific works and artists.

Religion. Essay questions that inquire about religious beliefs, doctrines, or practices are assigned this code.

Example:

Compare and contrast the Lutheran Reformation and the Catholic Reformation of the sixteenth century regarding the reform of both religious doctrines and religious practices.

The Renaissance

Example:

Explain the ways in which Italian Renaissance humanism transformed ideas about the individual's role in society.

The Enlightenment

Example:

"Napoleon was a child of the Enlightenment."

Assess the validity of the statement above. Use examples referring both to specific aspects of the Enlightenment and to Napoleon's policies and attitudes.

Women. Essay questions whose content involves women are assigned this code.

Example:

Compare and contrast the roles of British working women in the preindustrial economy (before 1750) with their roles in the era 1850 to 1920.

Everyday Life. Questions are assigned this code if they refer to household conditions, general working conditions, social patterns, education or other aspects of daily life, or to ideas and thinking about such topics. This code is also assigned to questions in which the words "social" or "cultural" refer to everyday life or to thinking or ideas about everyday life.

Example:

Analyze what differences in leisure activities shown in the two paintings on the preceding page reflect about the social life of peasants in the sixteenth century and of urban dwellers in the nineteenth century.

Appendix D: Content Categories for Biology Free-Response Questions

Listed below are content characteristics as they have been defined and operationalized. There are two sets of codes for the free-response questions. First, codes are assigned to the free-response questions themselves; then, codes are assigned separately to the scoring guides that are associated with the free-response questions.

Each code will be evaluated according to the following rating scale: 3 = dominant content; 2 = moderate content; 1 = weak content; 0 = no content. The free-response question itself will be assigned a 3 if the question requires that all or nearly all of the response should deal with this topic; in the case of the associated scoring guide, a 3 will be assigned if about 70 percent to 100 percent of the topics listed deal with this topic. The free-response question will be assigned a 2 if the code refers to one of a list of two or three topics that the question requires the AP candidate to write about; in the case of the associated scoring guide, if about 30 percent to 69 percent of the topics listed deal with this topic. The free-response question will be assigned a 1 if the code refers to one of a list of more than three topics the question requires the AP candidate to write about; in the case of the scoring guide, up to about 29 percent of the topics listed deal with this topic.

Content Hypothesized to Be Negatively Associated with Standardized Differences (favoring males)

Atmospheric science. The questions assigned this code deal with the earth's atmosphere.

Example:

Carbon is a very important element in living systems.

Explain how reactions involving carbon-containing compounds can contribute to the greenhouse effect.

Basic Chemistry. Essay questions assigned this code refer to basic chemistry of carbon and water, and concepts of chemistry, such as molarity.

Example:

The unique properties (characteristics) of water make life possible on Earth. Select three properties of water and:

For each property, identify and define the property and explain it in terms of the physical/chemical nature of water.

Applied Biochemistry. These questions deal with the application of biochemical knowledge and concepts.

Example:

Photosynthesis and cellular respiration recycle oxygen in ecosystems.

Explain how the metabolic processes of cellular respiration and photosynthesis recycle oxygen.

Enzymology. Questions assigned this code deal specifically with the biochemistry of enzymes.

Example:

Enzymes are biological catalysts.

Relate the chemical structure of an enzyme to its specificity and catalytic activity.

Experimental Apparatus. The questions assigned this code ask the student to design a piece of experimental apparatus.

Example:

The results below are measurements of cumulative oxygen consumption by germinating and dry seeds. Gas volume measurements were corrected for changes in temperature and pressure. [the table with results is not included here]

Describe the essential features of an experimental apparatus that could be used to measure oxygen consumption by a small organism.

Structure/Function Relationships. Questions assigned this code deal with biological structure/function relationships, ranging from biomolecules to organs.

Example:

Angiosperms (flowering plants) and vertebrates obtain nutrients from their environment in different ways.

Describe two structural adaptations in angiosperms for obtaining nutrients from the environment. Relate structure to function.

Math. Essay questions assigned this code require the student to construct a graph and/or to perform a calculation.

Example:

The results below are measurements of cumulative oxygen consumption by germinating and dry seeds. Gas volume measurements were corrected for changes in temperature and pressure. [the table with results is not included here]

a. Using the graph paper provided, plot the results for the germinating seeds at 22 deg. C and at 10 deg. C.

b. Calculate the rate of oxygen consumption for the germinating seeds at 22 deg. C, using the time interval between 10 and 20 minutes.

Content Hypothesized to Be Positively Associated with Standardized Differences (favoring females)

Cell Division. The questions assigned this code deal with both the meiotic and mitotic division of cells.

Example:

An organism is heterozygous at two genetic loci on different chromosomes. [the diagram associated with this question is not included here]

Explain how these alleles are transmitted by the process of mitosis to daughter cells.

Experimental Design. Questions assigned this code require the students to design an experiment.

Example:

Many physiological changes occur during exercise.

Design a controlled experiment to test the hypothesis that an exercise session causes short-term increases in heart rate and breathing rate in humans.

Genetics and Inheritance. Questions assigned this code refer to genetics and inheritance, including population genetics.

Example:

Assume that a particular genetic condition in a mammalian species causes an inability to digest starch. This disorder occurs with equal frequency in males and females. In most cases, neither parent of affected offspring has the condition.

Describe the most probable pattern of inheritance for this condition. Explain your reasoning. Included in your discussion a sample cross(es) sufficient to verify your proposed pattern.

Human Physiology. These questions deal specifically with the physiology of humans.

Example:

Describe negative and positive feedback loops, and discuss how feedback mechanisms regulate each of the following.

a. The menstrual cycle in a nonpregnant human female

b. Blood glucose levels in humans

Zoology/Classification. These questions deal with the evolutionary relationships between animals.

Example:

Describe the differences between the terms in each of the following pairs

1. Coelomate versus acoelomate body plan
2. Protostome versus deuterostome development
3. Radial versus bilateral symmetry

Appendix E: Content Categories for Microeconomics Free-Response Questions

Two categories of codes have been identified. Codes 01 through 05 identify processes involved in correctly answering the question while codes 06, 09, 10, and 12 relate to the subject matter addressed in the question. Codes have been assigned to the free-response questions themselves and then separately assigned to the scoring guides associated with the free-response questions. Especially in recent years the Advanced Placement economics questions have been broken into parts, and the scoring guides have been structured to correspond to those parts; therefore, marked similarities between the coding of the questions and the scoring guides can be expected to occur.

Each code will be evaluated according to the following rating scale: 3 = dominant content; 2 = moderate content; 1 = weak content; 0 = no content. The free-response question itself will be assigned a 3 if the question requires that all or nearly all of the response should deal with this process/topic; in the case of the associated scoring guide, a 3 will be assigned if about 70 percent to 100 percent of the point value of the question deals with this process/topic. The free-response question will be assigned a 2 if the code refers to one of a list of two or three processes/topics that the question requires the AP candidate to write about; in the case of the associated scoring guide, if about 30 percent to 69 percent of the point value of the question deals with this process/topic. The free-response question will be assigned a 1 if the code refers to one of a list of more than three processes/topics the question requires the AP candidate to write about; in the case of the scoring guide, up to about 29 percent of the point value of the question deals with this process/topic.

Content Hypothesized to Be Negatively Associated with Standardized Differences (favoring males)

Numerical/mathematical manipulation. This code is assigned to questions that involve mathematical calculations, reasoning, or manipulation of equations; understanding of mathematical relationships would also be included in this category.

Example:

The table above describes the production function for John Jones's T-shirt firm. Jones can hire as many workers as he wants for $75 per day and can sell as many T-shirts as he wants for $5 each. [table is not included here]

c. Assume the wage rate at which Jones can hire all the workers he wants increases to $120 per day, and the selling price of T-shirts increases to $6. Do each of the following.

(i) Explain how the demand for workers will change.

(ii) Indicate how many workers Jones will hire.

(iii) Indicate the quantity of T-shirts Jones will produce.

Graphical analysis. This code is assigned to questions in which the generation and/or interpretation of graphs is required or helpful. The category more strongly favors males when the graphical interpretation goes beyond simple supply and demand curve manipulation.

Example:

The demand and supply curves of cigarettes are depicted in the diagram above. [diagram is not included here]

Use supply and demand analysis to describe the impact of a per-unit tax on each of the following.

(i) The price paid by consumers for cigarettes

(ii) The quantity of cigarettes sold

Cause and effect reasoning. This code is assigned to questions requiring an understanding of causal relationships; multistep reasoning/analysis is required.

Example:

A perfectly competitive industry is in long-run equilibrium. The demand for the industry's product increases. Explain what will happen to the industry's output and price and to the typical firm's output and profit both in the short run and in the long run. Be sure to explain why the predicted outcomes will occur.

Factor market. This code is assigned if the subject matter deals with resource/factor markets and requires knowledge beyond simple supply and demand manipulation, i.e., how to determine the demand curve for a factor of production.

Example:

Initially a country's labor market is competitive and in long-run equilibrium. Now assume that new workers enter the labor market.

Assuming no other changes, explain how the increase in the number of workers will affect each of the following in the short run.

(i) The wage rate of workers

(ii) The price of goods produced by the workers

Price ceilings and floors. This code applies to questions relating to the effects of prices determined outside the market.

Example:

In a perfectly competitive market in long-run equilibrium, what would be the immediate results of imposing and enforcing a price ceiling below the equilibrium price of the product? What would be the long-run effect of continuing to enforce the ceiling price, assuming black markets do not develop? Be sure to explain why the predicted effects will occur.

Short-run and long-run effects. This code applies to questions that require students to differentiate between short-run and long-run effects of a given occurrence.

Example:

A perfectly competitive manufacturing industry is in long-run equilibrium. Energy is an important variable input in the production process and therefore the price of energy is a variable cost. The price of energy decreases for all firms in the industry.

a. Explain how and why the decrease in this input price will affect this manufacturing industry's output and price in the short run.

b. What will be the short-run effect on price, output, and profit of a typical firm in this manufacturing industry? Explain.

c. Will firms enter or exit this manufacturing industry in the long run? Why or why not?

Content Hypothesized to Be Positively Associated with Standardized Differences (favoring females)

Recognition and recall. This code is assigned to questions that can be answered with simple recall of facts, characteristics, definitions, and equations that are easily memorized, such as profit-maximization by producing that quantity where marginal revenue = marginal cost. This code is also assigned to questions the content of which has appeared on previously administered economics tests.

Example:

In the country of Lola, sugar had always been produced in a perfectly competitive industry until a dictator seized power and monopolized the production of sugar.

(A question like this has appeared in several previous tests.) Draw a graph that shows the output and price the monopolist would choose to maximize profits.

Application of a single concept. This code is assigned to questions that require a student to use an economic principle/concept to analyze a given situation. Only one concept or a one-step analysis is required.

Example:

Two goods, coffee and cream, are complements. Due to a natural disaster in Brazil that drastically reduces the supply of coffee in the world market, the price of coffee increases. Explain the effect of this increase in the price of coffee on each of the following.

a. The equilibrium price and quantity sold of cream.

b. The supply and demand for workers who produce cream.

Externalities. This code is assigned to content that relates to market failures.

Example:

Marginal analysis is essential to microeconomics decision making. Discuss how marginal analysis is used in each of the following cases.

To regulate an industry that produces a product that generates negative externalities.

Appendix F: Content Categories for Macroeconomics Free-Response Questions

Two categories of codes have been identified. Codes 01 through 05 identify processes involved in correctly answering the question while codes 06, 08, 10, 11, and 12 relate to the subject matter addressed in the question. Codes have been assigned to the free-response questions themselves and then separately assigned to the scoring guides associated with the free-response questions. Especially in recent years the Advanced Placement economics questions have been broken into parts, and the scoring guides have been structured to correspond to those parts; therefore, marked similarities between the coding of the questions and the scoring guides can be expected to occur.

Each code will be evaluated according to the following rating scale: 3 = dominant content; 2 = moderate content; 1 = weak content; 0 = no content. The free-response question itself will be assigned a 3 if the question requires that all or nearly all of the response should deal with this process/topic; in the case of the associated scoring guide, a 3 will be assigned if about 70 percent to 100 percent of the point value of the question deals with this process/topic. The free-response question will be assigned a 2 if the code refers to one of a list of two or three processes/topics that the question requires the AP candidate to write about; in the case of the associated scoring guide, if about 30 percent to 69 percent of the point value of the question deals with this process/topic. The free-response question will be assigned a 1 if the code refers to one of a list of more than three processes/topics the question requires the AP candidate to write about; in the case of the scoring guide, up to about 29 percent of the point value of the question deals with this process/topic.

Content Hypothesized to Be Negatively Associated with Standardized Differences (favoring males)

Numerical/mathematical manipulation. This code is assigned to questions that require mathematical calculation or reasoning or the manipulation of equations; understanding of mathematical relationships is also included in this category.

Example:

A stranger arrives from outside a given economic system with $1,000 of acceptable currency that has never been in the system before. The nation's banking system is governed by a central bank that has set a reserve requirement of 10 percent.

Assume the stranger deposited the $1,000 in a local bank. Explain the impact of this deposit on each of the following.

(i) The change in the dollar value of the local bank's reserves

(ii) The maximum possible change in the dollar value of the local bank's loans

(iii) The maximum possible change in the dollar value of the total money supply

Graphic analysis. This code is assigned to questions in which the generation and/or interpretation of graphs is required or helpful. The category more strongly favors males if the graphic analysis goes beyond simple aggregate supply and aggregate demand manipulation.

Example:

Assume that the Federal Reserve System sells bonds in the open market, and that commercial banks hold no excess reserves.

Show graphically what happens in the money market when the Federal Reserve sells bonds.

Cause and effect reasoning. This code is assigned to questions requiring an understanding of causal relationships. The student must be able to explain the transmission mechanisms by which a given action causes a particular result. Multistep reasoning/analysis is required.

Example:

Over the past two years, the unemployment rate in Country X has risen from five percent to nine percent. As the leader of Country X, you have been presented with two policy options to address the unemployment problem.

Policy 1: Use tariffs and quotas to restrict imports and thus protect jobs in Country X.

Policy 2: Use monetary and fiscal policies to solve the unemployment problem without resorting to trade restrictions.

Explain two disadvantages of selecting Policy 1.

Money, banking, and interest rates. This code is assigned to questions relating to money creation, the Federal Reserve System, and the effects of changes in the money supply on interest rates and economic performance in the short and long run.

Example:

Assume that the Federal Reserve System sells bonds in the open market, and that commercial banks hold no excess reserves.

Explain in detail how the Federal Reserve's action affects the commercial banks' reserves and the interest rates.

Unanticipated inflation. This code is assigned to questions testing the student's knowledge of the effects of unanticipated inflation on different groups of people.

Example:

Explain how some individuals are helped and others harmed by unanticipated inflation as they participate in each of the following markets.

a. Credit markets

b. Labor markets

c. Product markets

Content Hypothesized to Be Positively Associated with Standardized Differences (favoring females)

Recognition and recall. This code is assigned to questions that can be correctly answered with simple recall of facts, characteristics, definitions, and equations that are easily memorized. Examples of material in this category include AD = C+I+G+X; determinants of aggregate supply and aggregate demand; and the basic tenets of Keynesian theory. This code is also assigned to questions the content and/or format of which has appeared on previously administered AP Economics free-response tests.

Example:

The economy is at full employment. An increase in government spending causes the government deficit to increase.

Define each of the following.

(i) Government deficit

(ii) National debt

Application of a single concept. This code is assigned to questions that require a student to use an economic principle/concept to analyze a given situation. Only one concept, a one-step analysis, or application of several sequential single-step analyses is required.

Example:

The economy is at full employment. An increase in government spending causes the government deficit to increase.

Draw an aggregate supply and demand graph showing the economy at full employment. Show on the graph and explain completely the impact of the increase in government spending on each of the following.

(i) Price level

(ii) Real output

Effects of nominal wages and labor productivity. This code is assigned to questions relating to the effects of changes in nominal wages and/or productivity.

Example:

Assume that in the United States, nominal wage rates rise faster than labor productivity. Analyze the short-run effects of this situation on each of the following.

a. The general price level

b. The level of exports

c. The international value of the dollar

Aggregate demand, consumption, and the household sector. This code is assigned to questions that relate to determinants of aggregate demand and more especially to consumption.

Example:

Assume a market economy with a business sector, a household sector, and a government sector, but no international sector.

(a) Draw and label a circular flow diagram for this economy.

(b) Referring to the diagram you have drawn in part (a), identify two ways of calculating this economy's gross domestic product (GDP).

(c) Identify each of the following.

(i) The components of aggregate demand

(ii) The determinants of aggregate supply

Keynesian theory. This code is assigned to questions that require that the student understand Keynesian theory and the use of fiscal policy in particular to alleviate a recession.

Example:

Assume that the economy is in a recession.

a. Explain each of the following.

(i) Monetary and fiscal policies advocated by monetarists to eliminate the recession

(ii) Monetary and fiscal policies advocated by Keynesians to eliminate the recession

b. Explain how monetarists and Keynesians differ in their conclusions about the effects of crowding out associated with the stabilization policies outlined in Part (a).

www.collegeboard.com