Graduate Record Examinations®

Sex, Race, Ethnicity, and Performance on the GRE® General Test 1998-99

This report is a companion to the *GRE Guide to the Use of Scores.*

A Technical Report

Visit GRE Online http://www.gre.org

Published for the Graduate Record Examinations Board by Educational Testing Service

This report can be downloaded from the GRE Web site at www.gre.org.

Sex, Race, Ethnicity, and Performance on the GRE® General Test

A Technical Report

The concern for fairness pervades all aspects of testing including (1) the development of the tests, (2) the standardization of testing conditions, and (3) the use of the scores. Analysis of Graduate Record Examinations® (GRE®) General Test data reveals differences in the mean scores achieved by different racial, ethnic, and sex groups. Given this differential performance, questions have been raised about the possibility of intrinsic bias in the GRE General Test that could adversely affect women and minority test takers. However, test results cannot be judged in isolation from the unequal outcomes produced by our educational, economic, and social systems. A fair and accurate test mirrors real differences in relevant educational preparation. A test that did not reflect these differences would be an invalid indicator of educational accomplishment for all test takers.

Educational Testing Service (ETS) and the GRE Program have taken steps to ensure, to the extent possible, that tests and test scores are fair for all test takers, regardless of group membership. The purpose of this report is to discuss (1) the observed GRE General Test mean score differences that exist, (2) the procedures that ETS follows to ensure that its tests are fair to all individuals regardless of group membership, (3) the GRE research on validity of the GRE General Test with respect to different sex, age, racial, and ethnic groups, and (4) the specific score-use guidelines that are especially important in the context of test fairness.

Summary of GRE General Test Mean Score Differences

Examination of GRE General Test score data reveals mean score differences by racial, ethnic, and sex groups. In Appendix A, selected tables of GRE General Test score information are presented. Table A.1 presents GRE General Test score information by citizenship status and sex for the 1996-97 testing year. These data show that men tend to have higher mean scores, particularly on the quantitative measure. The table also shows that mean scores for non-U.S. citizens are higher than those for U.S. citizens on the quantitative measure, lower on the verbal measure, and similar on the analytical measure.

In Table A.2, test takers who identified themselves as White, Asian/Pacific, and Other have higher mean scores on the verbal and analytical measures than do the other ethnic groups. Asian/Pacific American test takers have the highest mean score on the quantitative measure. In addition, within each ethnic group and measure, mean scores are higher for men than women with the exception of American Indian women on the analytical measure.

Tables A.3, A.4, and A.5 present mean GRE General Test scores by broad intended graduate major field and sex for the 1996-97 testing year. Tables A.6, A.7, and A.8 present mean GRE General Test scores by broad intended graduate major field and racial/ethnic group for the 1996-97 testing year. In these tables, mean scores vary considerably by graduate major field. Admission to graduate school is generally determined on a field-by-field basis, and GRE General Test score data presented in the tables show that performance differences by sex, race, and ethnicity tend to be smaller within a field than are total mean score differences.

Steps that ETS Takes to Ensure Fairness

ETS has designed two procedures intended to build fairness into the tests. These procedures are the fairness review process and the Differential Item Functioning (DIF) analysis. The purpose of the fairness review process is to ensure that tests reflect the multicultural nature of American society, and to screen out any material that might be offensive to major subgroups of test takers, such as those based on age, disability, ethnic group, race, or sex. The purpose of the DIF analysis is to identify any test questions on which members of a particular group of test takers perform differently than would be expected on the basis of their overall ability in the areas covered by the test.

Fairness Review. All questions in any ETS test (and all materials published by ETS) must pass a fairness review. This review is based on a set of written guidelines; each review is conducted by an ETS staff member specifically trained in the application of these guidelines. Any test question that does not pass the fairness review must be revised to comply with the guidelines; otherwise it must be removed from the test. The fairness review does not guarantee that women, minority group members, or disabled persons will perform well on the test, but it does guard against the possibility of distraction caused by language or content that might be found offensive. Appendix B provides a summary of the ETS fairness review process.

DIF Analysis. Differential item functioning occurs when people of approximately equal knowledge and skill in different groups perform in substantially different ways on a test question. Differential Item Functioning analysis is a statistical technique intended to identify test questions that are more difficult for members of one group than for members of some other group, even when the groups are of comparable overall ability. It is important to realize that DIF is not synonymous with bias. DIF may occur if a perfectly fair question happens to be measuring a skill that is not well-represented in the test as a whole.

Appendix C provides detailed descriptions of the calculations of the DIF statistics. Each DIF analysis involves a set of comparisons between a group of examinees that is the focus of the study (focal group) and the group with which they are compared (reference group). If the focal group is women, the reference group is men. If the focal group is a minority group, the reference group consists of White test takers.

The DIF analysis is based on a comparison between groups of test takers of the same overall ability, as determined by their performance on the test as a whole. A DIF statistic is computed for each test question, indicating the extent to which members of the focal group perform differently from members of the reference group *who have the same scores on the test as a whole*. On the basis of this type of analysis, any questions that members of the focal group miss substantially more often than members of the reference group are deleted from the test scores used to match the two groups on ability. Then the DIF analysis is repeated, to see if this “purified criterion” reveals any additional questions that are particularly difficult for members of the focal group. There are two points in the testing process where DIF analyses are performed. If the questions have been pretested on an adequate number of test takers from the focal group, a DIF analysis is performed before the questions are selected for the test. In this case, a question showing a large DIF value will not be included in the test, unless that particular question is considered essential for the test’s content coverage. DIF analyses are also performed after the test is administered but before the scores are reported. In this case, a question showing a large DIF value is subjected to a special review by subject-matter experts and testing experts. The question is then included in the final computation of the test scores *only* if the experts agree that the question is substantively correct, correctly written, and important to the measurement purpose of the test.

In 1994 ETS instituted a set of guidelines, based on many years of research related to DIF statistics, that identified several content categories of questions that sometimes produce negative DIF. The Corporation decided to prohibit for skills tests further use of questions in those categories, regardless of the performance with respect to DIF for particular questions in those categories.

Research on Differential Item Functioning

GRE research on DIF has shown that the majority of GRE General Test questions do not show large DIF statistics (96 percent of GRE General Test items from seven recent forms of the test did not show large DIF values). Specific DIF research studies are included in the bibliography.

Research on Validity

ETS and the GRE Program have conducted research on the relationship between GRE General Test scores and graduate school performance. Since the main use of the GRE scores is to predict academic success in graduate school, research has tended to focus on the relationship between GRE General Test scores and graduate school grades for different groups of graduate students. Although the sample size of minorities is not large enough to be definitive, the available data do not show evidence of bias. The data have consistently shown that the scores generally predict about as well for test takers of one sex as for the other, and about equally well for test takers of one racial or ethnic group as for another.

There are two exceptions to this general pattern of results: older students and students in the natural sciences whose best language is not English. When students over age 24 are considered as a separate group, GRE quantitative and analytical scores and undergraduate grades tend to slightly underestimate their graduate grades. This underestimation appears particularly true of women over 24, who on the average obtain graduate grades about two-tenths of a grade point higher than students 24 or under with the same undergraduate grades and GRE scores.

On the average, nonnative English speaking students in the natural sciences obtain higher graduate grades (by about one-tenth of a grade point) than native English speakers with the same GRE verbal scores. To a lesser extent this occurs with GRE analytical scores, but does not occur with GRE quantitative scores.

Implications for Test Use

Existing research shows GRE General Test scores underestimate the graduate grades of older students, particularly women, and GRE verbal scores underestimate the graduate grades of nonnative English speaking students in the natural sciences. This underprediction should be taken into account in selecting students for graduate programs. For example, in comparing applicants with similar GRE scores and undergraduate grades, programs could choose to accept a higher proportion of women over 24 years old compared to men and women 24 years old or younger. Data are not yet available to determine if similar underprediction also occurs for other important criteria of graduate school success.

Score Use Guidelines. The Guidelines for the Use of GRE Scores provides information about the appropriate use of GRE test scores. Because differences exist between GRE General Test mean scores of groups based on ethnicity, race, or sex, adherence to these guidelines is critical to ensure a fair graduate application process. The complete Guidelines for the Use of GRE Scores are included in the *GRE Guide to the Use of Scores*, which is a companion to this report. Four of the guidelines that are especially important in the context of test fairness are summarized below.

1. *Use of multiple criteria (in the admissions process).* No single measure — and this includes the GRE General Test — assesses every discipline-related skill necessary for academic work or all subjective factors important to academic and career success, such as motivation, creativity, and interpersonal skills. Therefore, all available pertinent information about an applicant should be considered when making a decision. In view of the breadth of information relevant to judging success in graduate education, the GRE Board believes it is inadvisable to reject an applicant solely on the basis of GRE scores.

2. *Consideration of verbal, quantitative, and analytical scores as three separate and independent measures*. An applicant with a 300 on the verbal measure and an 800 on the quantitative measure is very different from another applicant with an 800 on verbal and a 300 on quantitative. The former applicant might do well in a mathematics program, but the latter probably would not. Similarly, the student with the 800 on the verbal measure might have a high probability of success in an English literature program. Summing GRE verbal and quantitative scores hides the differences between these applicants. Further, summing the scores and then blindly applying a minimum combined score (cutoff score), such as V+Q must be greater than 1100, may eliminate qualified individuals from the applicant pool.

3. *Avoidance of decisions based on small score differences*. Because of psychometric limitations, only score differences of certain magnitudes are reliable indicators of real differences in performance. A person's test score is an estimate of the level of the person's knowledge or ability in the area tested and is not a complete and perfect measure. The standard error of measurement is an index of the variation in scores to be expected because of imprecise measurement. When the test scores of two test takers are compared, the difference between their scores will be affected by errors of measurement in each of the scores. Small differences in scores may be due to measurement error and not to differences in the abilities of the test takers. The values of the standard error of measurement of score differences should be used when comparing the scores of test takers because small score differences may not represent real differences in the abilities of the test takers. Users of GRE test scores are thus cautioned in the *GRE Guide to the Use of Scores* not to make fine distinctions when comparing the scores of two or more test takers.

4. *Conducting validity studies*. Institutions using GRE scores in the admissions process are encouraged to examine the relationship between test scores and measures of performance in their academic programs. The GRE Program staff will provide advice on the design of appropriate validation studies without charge. Information on further validation procedures can be obtained from the technical standards on testing of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education.

Conclusion

Issues of fairness are a constant concern to the developers of the GRE tests. One type of concern involves the content, the wording, and the statistical characteristics of individual test questions. Another type involves the relationship between GRE scores and graduate school grades. ETS addresses the first type of concern by using procedures designed to exclude from the tests any questions that might tend to make the tests unfair to women or to members of racial or ethnic minority groups. ETS addresses the second type of concern by conducting and publishing statistical studies of the relationships between GRE scores and grades for various groups of test takers and publishing guidelines for appropriate use of the scores. Taken as a whole, the Guidelines for the Use of GRE Scores promotes fairness by pointing out the limitations of test scores and the need for flexibility in their use. Tests, however, are only fair or unfair in the context of how they are used. Thus, the GRE Program and the users of GRE scores both have a responsibility for assuring test use that is not discriminatory on the basis of sex, race, or ethnicity.

BIBLIOGRAPHY

Allen, N., and Wainer, H. 1989. Nonresponse in declared ethnicity and the identification of differentially functioning items. Princeton, New Jersey. Educational Testing Service, *Program Statistics Research Technical Report* No. 89-89.

Braun, H., and Jones, D. 1985. Use of empirical Bayes methods in the study of the validity of academic predictors of graduate school performance. Princeton, New Jersey. Educational Testing Service, *Research Report* No. 84-34.

Donlon, T. F.; Hicks, M. M.; and Wallmark, M. M. 1980. Sex differences in item responses on the Graduate Record Examinations. *Applied Psychological Measurement*, 1980, 4, pp. 9-20.

Educational Testing Service 1992. *ETS sensitivity review process*. Princeton, NJ: Educational Testing Service.

Educational Testing Service 1995. *Guidelines for the use of GRE scores*. Princeton, NJ: Educational Testing Service.

Freedle, R., and Kostin, I. 1988. Relationship between item characteristics and an index of differential item functioning for the four GRE verbal item types. Princeton, New Jersey: *GRE Board Professional Report* No. 85-3P, ETS Research Report No. 88-29.

Grandy, J. 1994. *GRE Trends and Profiles: Statistics about General Test examinees by sex and ethnicity.* Princeton, NJ: Educational Testing Service.

Holland, P. W., and Thayer, D. R. 1986. Differential item performance and the Mantel-Haenszel procedure. Paper presented at the annual meeting of the *American Educational Research Association*, San Francisco, California, April 1986.

Kingston, N. M.; Schneider, L. M.; and Briel, J. B. 1988. Using empirical Bayes methods to investigate the potential sex and ethnic bias of the GRE General Test. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C., August 1988.

McPeek, W. M., and Wild, C. L. 1986. Performance of the Mantel-Haenszel statistic in a variety of situations. Paper presented at the annual meeting of the *American Educational Research Association*, San Francisco, California, April 1986.

McPeek, W. M., and Wild, C. L. 1987. Characteristics of quantitative items that function differently for men and women. Paper presented at the annual meeting of the *American Psychological Association*, New York, New York, August 1987.

Scheuneman, J. D. 1985. Exploration of causes of bias in test items. Princeton, New Jersey: *Graduate Record Examinations Board Professional Report* GREB No. 81-21P, ETS Research Report No. 85-42.

Stricker, L. J. 1981. A new index of differential subgroup performance: Application to the GRE Aptitude Test. Princeton, New Jersey: *Graduate Record Examinations Board Professional Report* GREB No. 78-7P. ETS Research Report No. 81-13.

Stricker, L. J., and Rock, D. A. 1985. Factor structure of the GRE General Test for older examinees: Implications for construct validity. Princeton, New Jersey: *Graduate Record Examinations Board Professional Report* GREB No. 83-10R, ETS Research Report No. 85-9.

Swinton, S. S. 1987. The predictive validity of the restructured GRE with particular attention to older students. Princeton, New Jersey: *Graduate Record Examinations Board Professional Report* No. 83-25P, ETS Research Report No. 87-22.

Wah, D., and Robinson, D. 1990. *Examinee and score trends for the GRE General Test: 1977-78, 1982-83, 1986-87, and 1987-88.* Princeton, NJ: Educational Testing Service.

Wild, C.; McPeek, W. M.; and Koffler, S. L. 1988. Concurrent validity of verbal item types for ethnic and gender subgroups. Princeton, New Jersey. Educational Testing Service, *Graduate Record Examinations Board Technical Report* GRE No. 84-10.

Willingham, W. W., and Cole, N.S. 1997. *Gender and Fair Assessment*, Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Willingham, W. W., and Johnson, L. M. (Eds.) 1997. *Supplement to Gender and Fair Assessment.* Princeton, NJ: Educational Testing Service, Research Report No. 97-1.

APPENDIXES

APPENDIX A

Table A.1

GRE General Test Score Information by Citizenship Status and Sex: 1996-97

	Examinees		GRE General Test Score Information					
			Verbal		Quantitative		Analytical	
Group	Number	Percent*	Mean	SD	Mean	SD	Mean	SD
U.S. Citizens	277,523	76	483	108	529	127	547	127
Men	98,314	27	500	110	576	128	563	130
Women	179,209	49	474	106	504	119	539	125
Non-U.S. Citizens	87,010	24	426	119	649	129	539	133
Men	51,261	14	427	120	675	116	544	135
Women	35,749	10	424	119	612	137	532	131
Total	364,606	100	470	113	558	137	545	129
Men	149,608	41	475	119	610	133	556	132
Women	214,998	59	465	109	522	128	538	126

Note: A total of 394,396 examinees took the GRE General Test in 1996-97 and 92 percent responded to questions in this table.

*Percentages in this table are based on the column total.

Table A.2

GRE General Test Score Information by Ethnic Group and Sex: 1996-97
(U.S. Citizens Only)

			GRE General Test Score Information					
	Examinees		Verbal		Quantitative		Analytical	
Group[1]	Number	Percent[2]	Mean	SD	Mean	SD	Mean	SD
American Indian	1,586	1	456	99	479	118	506	124
Men	588	<1	466	101	511	123	502	126
Women	998	<1	451	97	459	110	508	124
Asian/Pacific	13,115	5	487	117	598	124	557	127
Men	5,097	2	490	121	638	117	566	131
Women	8,018	3	484	114	572	121	552	124
Black/African	22,923	8	391	90	416	112	423	111
Men	6,410	2	399	93	446	125	429	118
Women	16,513	6	388	88	404	104	421	109
Mexican American	5,755	2	435	97	475	123	483	122
Men	2,120	1	449	101	517	131	495	126
Women	3,635	1	427	94	451	112	476	118
Puerto Rican	2,837	1	410	100	471	121	459	121
Men	1,148	<1	414	100	504	128	462	125
Women	1,689	1	408	99	448	111	457	118
Other Hispanic	5,539	2	455	104	495	125	499	128
Men	1,923	1	472	108	543	127	514	131
Women	3,616	1	445	101	469	116	490	126
White	215,423	79	495	103	540	121	564	120
Men	76,441	28	512	105	586	121	578	122
Women	138,982	51	485	101	514	113	556	118
Other	7,118	3	510	118	541	131	555	132
Men	3,119	1	527	118	583	129	572	133
Women	3,999	2	497	116	509	123	541	130
Total	274,296	100	483	108	529	127	547	127
Men	96,846	35	500	110	576	128	563	130
Women	177,450	65	473	105	503	118	539	125

Note: A total of 277,523 U.S. citizens took the GRE General Test in 1996-97 and 99 percent responded to questions in this table.

[1] Ethnic groups are defined as follows: American Indian–American Indian, Inuit, or Aleut; Asian/Pacific American–Asian or Pacific American; Black/African American–Black or African American; Mexican American–Mexican American or Chicano; Puerto Rican–same; Other Hispanic–Other Hispanic or Latin American; White–same; Other–same.

[2] Percentages in this table are based on the column total and are rounded to the nearest integer.

Table A.3

Mean GRE General Test Verbal Scores by Intended Graduate Major Field and Sex: 1996-97

Graduate Major		Men	Women	No Response	Total
Business	N	5,354	6,559	39	11,952
	Mean	448	437	454	442
	SD	107	105	104	106
Education	N	6,040	21,594	95	27,729
	Mean	428	426	447	426
	SD	94	90	96	91
Engineering	N	32,810	8,385	73	41,268
	Mean	453	460	429	455
	SD	120	118	128	120
Humanities and Arts	N	18,078	30,844	146	49,068
	Mean	540	517	501	526
	SD	118	118	127	118
Life Science	N	24,453	54,964	184	79,601
	Mean	468	458	441	461
	SD	104	98	107	100
Physical Science	N	19,513	12,042	64	31,619
	Mean	484	471	439	479
	SD	124	119	116	123
Social Science	N	24,543	52,303	202	77,048
	Mean	487	471	471	476
	SD	113	105	113	108
Other Fields	N	9,963	20,406	117	30,486
	Mean	463	448	444	453
	SD	113	106	108	109
No Response	N	21,465	21,672	2,488	45,625
	Mean	467	459	459	463
	SD	122	117	115	119
Total	N	162,219	228,769	3,408	394,396
	Mean	475	465	459	469
	SD	119	109	115	113

Table A.4

Mean GRE General Test Quantitative Scores by Intended Graduate Major Field and Sex: 1996-97

Graduate Major		Men	Women	No Response	Total
Business	N	5,354	6,559	39	11,952
	Mean	552	510	496	529
	SD	132	133	140	134
Education	N	6,040	21,594	95	27,729
	Mean	498	458	446	466
	SD	120	109	113	113
Engineering	N	32,810	8,385	73	41,268
	Mean	697	683	685	694
	SD	88	93	77	90
Humanities and Arts	N	18,078	30,844	146	49,068
	Mean	552	515	490	529
	SD	125	119	125	122
Life Science	N	24,453	54,964	184	79,601
	Mean	591	529	507	548
	SD	116	118	131	121
Physical Science	N	19,513	12,042	64	31,619
	Mean	683	638	622	666
	SD	103	115	104	110
Social Science	N	24,543	52,303	202	77,048
	Mean	559	506	500	523
	SD	129	118	126	124
Other Fields	N	9,963	20,406	117	30,486
	Mean	539	482	474	501
	SD	129	121	118	127
No Response	N	21,465	21,672	2,488	45,625
	Mean	644	561	522	598
	SD	140	154	142	153
Total	N	162,219	228,769	3,408	394,396
	Mean	613	525	520	561
	SD	134	131	140	139

Table A.5

Mean GRE General Test Analytical Scores by Intended Graduate Major Field and Sex: 1996-97

Graduate Major		Men	Women	No Response	Total
Business	N	5,354	6,559	39	11,952
	Mean	515	512	482	514
	SD	137	133	134	135
Education	N	6,040	21,594	95	27,729
	Mean	487	487	471	487
	SD	128	122	110	123
Engineering	N	32,810	8,385	73	41,268
	Mean	572	595	547	577
	SD	134	124	119	132
Humanities and Arts . .	N	18,078	30,844	146	49,068
	Mean	565	552	516	557
	SD	132	126	140	129
Life Science	N	24,453	54,964	184	79,601
	Mean	551	541	497	544
	SD	128	123	129	125
Physical Science	N	19,513	12,042	64	31,619
	Mean	593	587	547	591
	SD	133	127	145	131
Social Science	N	24,543	52,303	202	77,048
	Mean	546	541	510	542
	SD	133	125	140	128
Other Fields	N	9,963	20,406	117	30,486
	Mean	522	513	468	516
	SD	134	127	126	130
No Response	N	21,465	21,672	2,488	45,625
	Mean	556	535	510	543
	SD	140	135	143	138
Total	N	162,219	228,769	3,408	394,396
	Mean	556	538	508	545
	SD	135	128	141	132

Table A.6

Mean GRE General Test Verbal Scores by Intended Graduate Major Field and Ethnic Group: 1996-97
(U.S. Citizens Only)

Graduate Major		American Indian	Asian/ Pacific American	Black/ African American	Mexican-American	Puerto Rican	Other Hispanic Latin-American	White	Other	No Response	Total
Business	N	48	269	1,500	225	104	187	6,568	157	53	9,132
	Mean	464	445	369	405	393	440	472	450	445	451
	SD	114	109	78	86	84	103	95	112	118	101
Education	N	148	349	2,149	663	146	508	20,593	317	125	25,062
	Mean	422	418	359	384	388	393	439	425	447	429
	SD	86	89	71	77	86	75	87	101	107	89
Engineering	N	57	1,851	1,003	352	329	381	10,492	360	152	14,992
	Mean	471	478	427	452	394	472	521	517	548	504
	SD	85	121	95	92	86	106	95	115	111	105
Humanities and Arts ...	N	189	1,373	2,153	778	357	778	32,732	1,423	500	40,364
	Mean	512	542	437	476	464	497	550	556	577	541
	SD	105	112	104	111	114	118	104	111	106	109
Life Science	N	408	3,499	3,952	989	655	1,012	51,130	1,158	401	63,299
	Mean	449	481	389	436	398	449	475	490	502	468
	SD	90	109	80	91	92	94	93	108	107	96
Physical Science	N	77	1,091	1,495	242	210	267	13,952	481	213	18,047
	Mean	491	485	399	454	396	487	526	541	568	511
	SD	105	132	88	107	115	109	103	120	117	112
Social Science	N	395	2,855	6,256	1,625	626	1,565	47,940	1,808	554	63,748
	Mean	455	490	394	440	422	458	495	503	521	482
	SD	98	110	91	93	94	100	101	111	115	105
Other Fields	N	148	867	2,640	532	219	513	18,823	644	177	24,629
	Mean	434	466	378	428	401	435	476	492	498	463
	SD	89	116	84	94	99	95	101	119	117	105
No Response	N	116	961	1,775	349	191	328	13,193	770	1,123	18,920
	Mean	435	493	375	417	392	442	501	512	501	485
	SD	96	122	89	96	99	106	109	128	118	116
Total	N	1,586	13,115	22,923	5,755	2,837	5,539	215,423	7,118	3,298	278,193
	Mean	456	487	391	435	410	455	495	510	520	483
	SD	99	117	90	97	100	104	103	118	118	108

Table A.7

Mean GRE General Test Quantitative Scores by Intended Graduate Major Field and Ethnic Group: 1996-97
(U.S. Citizens Only)

Graduate Major		American Indian	Asian/ Pacific American	Black/ African American	Mexican-American	Puerto Rican	Other Hispanic Latin-American	White	Other	No Response	Total
Business	N	48	269	1,500	225	104	187	6,568	157	53	9,132
	Mean	473	549	394	449	450	476	534	491	507	506
	SD	114	122	94	109	117	119	110	130	149	121
Education	N	148	349	2,149	663	146	508	20,593	317	125	25,062
	Mean	433	498	373	406	410	429	476	447	456	463
	SD	96	116	87	96	89	100	106	114	130	109
Engineering	N	57	1,851	1,003	352	329	381	10,492	360	152	14,992
	Mean	636	693	566	633	593	630	691	676	706	678
	SD	94	84	118	92	97	97	78	100	86	91
Humanities and Arts	N	189	1,373	2,153	778	357	778	32,732	1,423	500	40,364
	Mean	471	572	415	464	446	477	531	531	557	523
	SD	117	113	106	116	119	120	113	120	121	117
Life Science	N	408	3,499	3,952	989	655	1,012	51,130	1,158	401	63,299
	Mean	488	596	426	498	467	513	540	551	575	534
	SD	108	115	102	113	108	119	109	120	122	115
Physical Science	N	77	1,091	1,495	242	210	267	13,952	481	213	18,047
	Mean	586	662	513	610	534	608	655	652	674	641
	SD	120	107	118	110	121	114	100	105	107	111
Social Science	N	395	2,855	6,256	1,625	626	1,565	47,940	1,808	554	63,748
	Mean	475	559	404	462	444	482	525	521	546	511
	SD	114	120	102	110	109	113	112	122	135	118
Other Fields	N	148	867	2,640	532	219	513	18,823	644	177	24,629
	Mean	442	549	380	439	436	458	500	502	523	486
	SD	100	124	94	112	112	110	113	127	132	119
No Response	N	116	961	1,775	349	191	328	13,193	770	1,123	18,920
	Mean	428	592	390	444	439	464	531	541	536	517
	SD	112	129	109	123	116	130	123	136	131	132
Total	N	1,586	13,115	22,923	5,755	2,837	5,539	215,423	7,118	3,298	278,193
	Mean	479	598	416	475	471	495	540	541	558	529
	SD	118	124	112	123	121	125	121	131	136	127

Table A.8

Mean GRE General Test Analytical Scores by Intended Graduate Major Field and Ethnic Group: 1996-97
(U.S. Citizens Only)

Graduate Major		American Indian	Asian/ Pacific American	Black/ African American	Mexican- American	Puerto Rican	Other Hispanic Latin- American	White	Other	No Response	Total
Business	N	48	269	1,500	225	104	187	6,568	157	53	9,132
	Mean	502	511	401	448	434	468	546	495	502	514
	SD	123	132	100	116	123	121	118	139	143	129
Education	N	148	349	2,149	663	146	508	20,593	317	125	25,062
	Mean	460	493	384	419	402	436	507	464	479	491
	SD	110	119	91	105	105	108	114	122	133	119
Engineering	N	57	1,851	1,003	352	329	381	10,492	360	152	14,992
	Mean	562	585	497	540	496	552	635	602	638	611
	SD	130	129	123	117	110	124	108	134	124	122
Humanities and Arts	N	189	1,373	2,153	778	357	778	32,732	1,423	500	40,364
	Mean	525	571	442	499	478	509	579	577	603	568
	SD	127	119	116	127	131	131	116	125	117	122
Life Science	N	408	3,499	3,952	989	655	1,012	51,130	1,158	401	63,299
	Mean	509	556	428	496	451	511	561	550	572	549
	SD	119	125	106	118	113	127	115	127	125	120
Physical Science	N	77	1,091	1,495	242	210	267	13,952	481	213	18,047
	Mean	566	581	467	542	458	557	625	619	655	605
	SD	114	131	119	128	132	127	111	122	120	124
Social Science	N	395	2,855	6,256	1,625	626	1,565	47,940	1,808	554	63,748
	Mean	518	550	427	491	470	501	565	550	574	545
	SD	124	122	111	115	120	125	117	127	133	125
Other Fields	N	148	867	2,640	532	219	513	18,823	644	177	24,629
	Mean	475	532	403	472	450	479	540	531	540	521
	SD	116	127	103	119	118	127	120	137	145	127
No Response	N	116	961	1,775	349	191	328	13,193	770	1,123	18,920
	Mean	456	545	392	441	421	463	546	541	539	526
	SD	130	132	107	117	114	128	122	137	130	131
Total	N	1,586	13,115	22,923	5,755	2,837	5,539	215,423	7,118	3,298	278,193
	Mean	506	557	423	483	459	499	564	555	568	547
	SD	124	127	111	122	121	128	120	132	134	127

Appendix B

The ETS Fairness Review Process (Excerpted from ETS Fairness Review Process: An Overview)

Reviewers

Reviews of ETS publications are conducted by ETS professional staff members who are trained in fairness issues at one-day workshops and periodic refresher courses. While there are a number of reviewers who are women and/or members of minority groups, membership in such groups is not a prerequisite, and any professional interested in the process and showing concern for equity may be trained to administer it.

Test Fairness Review Procedures

The test fairness review process has three components: an optional preliminary review (required by some testing programs), a mandatory final review, and an adjudication process.

1. *Preliminary review*. Any staff member who is assembling a test may request a preliminary review to screen questions and answers, reading passages, and other materials for sensitivity-related issues. The reviewer's recommendations are not binding at this stage; however, a preliminary review is an excellent means of identifying potential problems early in the test development process, when modifications can be made more easily.

2. *Final review*. The mandatory final review takes place after the test has been assembled and during the regular editorial process. This review must be conducted, even if the test received a preliminary review.

The fairness reviewer, who is always someone other than the person who is responsible for the test (the test assembler), notifies the test assembler in writing of any sensitivity-related issues the test has raised. The test assembler must then address in writing all concerns of the sensitivity reviewer. In the vast majority of cases, the test assembler and the reviewer are able to resolve the issues satisfactorily. When the two cannot resolve issues raised by the reviewer, a fairness review coordinator meets with them to ensure that they clearly understand each other's position. If the reviewer and assembler still cannot reconcile their differences, the disagreement is submitted to a panel for adjudication.

3. *Adjudication*. Adjudication is performed by a panel of fairness review coordinators from the test assembler's and fairness reviewer's areas, the test assembler's group head, and the director of test development. After examining the disputed material, the panel attempts to reach consensus. If consensus cannot be reached, a binding decision is made by the test assembler's director of test development.

Review Criteria

The fairness review training sessions teach reviewers to evaluate material in light of specific criteria:

1. *Stereotyping*. All ETS publications are reviewed to ensure that their language and illustrations reflect a fair and unbiased attitude toward all people and are free of material that reinforces stereotypes.

2. *Examinee perspective*. Test fairness reviewers have a particular concern that does not apply often to reviewers of other kinds of publications. They must evaluate all questions from the perspective of test takers, who do not necessarily know the correct answers. If an examinee must know the correct answer in order to prevent a question from reinforcing negative attitudes or stereotypes, the question may be in violation of the guidelines.

3. *Underlying assumptions*. While stereotypes are often blatant, underlying assumptions can be extremely subtle. Underlying assumptions may lead one to mistake aspects of Western culture for universal norms or to misunderstand a particular group. For instance, a publication that refers to an "afflicted" person "suffering from" cerebral palsy reflects the writer's underlying assumptions about what it is like to have this physical condition.

4. *Controversial material*. Highly controversial material, such as legalized abortion, is to be included in tests only when it is relevant to what is being tested. For example, a test for doctors or nurses may have to contain questions on abortion, but a test of reading ability should not include a reading passage on this controversial subject.

5. *Contextual considerations*. Sometimes the use of potentially sensitive material is unavoidable. There are four main areas in which this may occur:

    - *Historical domain:* In order to measure an individual's knowledge of history, it may sometimes be necessary to quote from material written during a period when social values differed markedly from today's. For example, an older passage describing members of the Black community may use the term "colored." While it is desirable to avoid such material when possible, the material must be judged in the overall context in which it appears.
    - *Literary domain:* Material that is designed to measure an individual's knowledge of literature or quotes from works of literature often contains similar problems. For example, a passage may use the so-called "generic he" in referring to men and women. Again, such material must be evaluated in light of the overall purpose of the test.
    - *Legal domain:* Material drawn from legal sources may sometimes deal with sensitive issues. For example, a law test question on the detention of citizens may refer to the incarceration of Japanese Americans during World War II.
    - *Health domain:* Certain examinations in the health professions require knowledge that may be considered sensitive in other contexts. For example, it may be necessary to test nursing candidates' knowledge of Tay-Sachs disease in Jewish families.

    Inclusion of potentially sensitive material depends on the content of the entire test or publication. Given an appropriate context, use of certain material may be justifiable.

6. *Elitism, ethnocentricity, and related problems*. To eliminate concepts, words, phrases, or examples that may upset or otherwise disadvantage a test taker, ETS makes every effort not to include expressions that might be more familiar to members of a particular social class or ethnic group than the general population, such as "soul food" and "trust fund," unless the terms are defined or knowledge of them is relevant to the purpose of the test. Words and sentence constructions that could have different meanings for different ethnic or geographic groups are avoided. Care is also taken to assess the appropriateness of dialect, slang, and non-English words and phrases, such as "bairn," "stickball," and "maven," which tend to be more familiar to certain ethnic, geographic, or other subgroups of English speakers.

Appendix C

Differential Item Difficulty Statistics and Categories

Overview

This appendix provides more detailed descriptions of the calculations of the Mantel-Haenszel and Standardized P-Difference statistics and of the assignment of questions to categories than were provided in the body of the report. The descriptions of the calculations are designed for readers who are not specialists in statistics. Readers with training in statistics may prefer the level of detail to be found in the following publications:

Dorans, N. "Two new approaches to assessing differential item functioning: standardization and the Mantel-Haenszel method." *Applied Measurement in Education*, *2*, no. 3, 1989, pp. 217-233.

Holland, P. and Thayer, D. "Differential item performance and the Mantel-Haenszel procedure." In Wainer, H., and Braun, H. (Eds.) *Test Validity*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1988.

Mantel, N., and Haenszel, W. "Statistical aspects of the analysis of data from retrospective studies of disease." *Journal of the National Cancer Institute*, *22*, 1959, pp. 719-748.

The Mantel-Haenszel Statistic

In its use with tests, the Mantel-Haenszel statistic is based on a comparison of the odds of answering a question correctly for matched people in the groups being compared. In operational use of indices of differential item difficulty at ETS, people are matched on the basis of test scores or subscores. The scores have been shown to be reliable and valid, and they are obtained under standardized conditions for all examinees. Even though people with the same test scores are not identical, they are likely to be reasonably well matched in terms of the knowledge and skill measured by the test.

To help ensure that the test scores used for matching are themselves free of questions that may be biased, the DIF analysis at ETS is generally done in two steps. First, a preliminary analysis is completed and any Category C questions (see page 22) are removed from the test on which people are to be matched. The modified test scores are then used as the matching criteria for the production of the DIF statistics that are to be used operationally for all of the questions in the test.

If, for example, a test has 100 questions, three questions may be removed from the matching criterion as a result of the preliminary DIF analysis. On the basis of the remaining 97 questions, the people who have taken the test can be divided into as many as 98 clusters based on the number of questions they answered correctly: one cluster containing people with scores of zero, another cluster containing people with scores of one, and so on up to a final cluster containing people with scores of 97. Even though no test can be a perfect measure of any knowledge or skill, the people within each score level should be quite similar in terms of what the test is measuring.

The procedure looks within each cluster of people at a score level and calculates the odds that members of the two groups being compared will answer the question correctly. For example, if there are 20 women at a particular score level and 16 of them answer correctly, then the odds are 16/4 or 4 to 1 that a woman at that score level will answer correctly. If 12 out of 18 men answer the questions correctly, then the odds are 12/6 or 2 to 1 that a man at that score level will answer the question correctly.

The next step in the procedure is to calculate the ratio of the two odds to obtain an indication of the relative advantage of one group over the other within the score level. For our example, the ratio is 4/1 (the women's odds) divided by 2/1 (the men's odds), which equals 2. This indicates that the women's odds of answering the question correctly are twice as great as the men's odds for people in that particular score level. The "odds ratios" are then averaged across all of the score levels using statistically optimal weights. See Holland and Thayer (1988) for a fuller description of the weighting procedure.

The Mantel-Haenszel statistic can be defined as the average factor by which the odds that members of one group will answer a question correctly exceed the corresponding odds for *comparable* members of the other group. The Mantel-Haenszel statistic is, therefore, in the form of an odds ratio. To obtain a statistic that is more meaningful to ETS test developers, the odds ratios are transformed to an index that can be interpreted directly in terms of differences in the difficulty of questions. The DIF statistic is expressed as *differences* on the delta scale that is commonly used by test developers at ETS to indicate the difficulty of test questions.[1]

[1] The delta scale is an inverse normal transformation of percent correct to a linear scale with a mean of 13 and standard deviation of 4.

For that statistic, known as MH D-DIF, a value of 1.00 means that one of the two groups being analyzed found the question to be one delta point harder than did *comparable* members of the other group.

The Standardized P-Difference

The other DIF statistic in routine use at ETS is called the Standardized P-Difference. To compute this statistic, we first classify all the examinees in each of the two groups being compared according to their scores. At each score level, we compute the proportion of examinees answering the question correctly in each of the two groups being compared (male and female examinees, Black and White examinees, etc.). We compute the difference between these two proportions at each score level. Then we combine the data for all the score levels in the following way: (1) we multiply the difference between groups at each score level by the percent of the focal group scoring at that level; (2) we combine these weighted differences to get a weighted average difference. This weighted average difference between the two groups is the Standardized P-Difference. A concise way to describe this procedure is to say that we first compute the difference between groups separately at each score level, using all available focal group and reference group examinees, and then average the differences over all the score levels, using the frequency distribution of scores in the focal group as weights. Computing a weighted average with weights based on the relative frequency of scores in the focal group has the effect of emphasizing the differences at those score levels with the greatest concentration of focal group members.

Categories of Questions

The Mantel-Haenszel odds ratios are transformed to differences on the delta scale as described in the referenced work by Holland and Thayer. The resulting statistics, MH D-DIF, are used to assign questions to categories (MH DIF CLASS) according to the following rules:

MH DIF CLASS A) MH D-DIF *not* significantly different from zero
OR
absolute value less than 1.0

MH DIF CLASS B) MH D-DIF significantly different from zero and absolute value of at least 1.0
AND EITHER
(1) less than 1.5
OR
(2) not significantly greater than 1.0

MH DIF CLASS C) MH D-DIF significantly greater than 1.0
AND
absolute value 1.5 or more.

For further clarification, note that the category can be determined by answering four questions in sequence:

1. Is the absolute value of MH D-DIF significantly greater than zero?

   No. Question is MH DIF CLASS A.

   Yes. Go on.

2. Is the absolute value of MH D-DIF 1.0 or more?

   No. Question is MH DIF CLASS A.

   Yes. Go on.

3. Is the absolute value of MH D-DIF significantly greater than 1.0?

   No. Question is MH DIF CLASS B.

   Yes. Go on.

4. Is the absolute value of MH D-DIF 1.5 or more?

   No. Question is MH DIF CLASS B.

   Yes. Question is MH DIF CLASS C.

50% RECYCLED PAPER
10% Post Consumer Waste

54053-06387 • C88M35• Printed in U.S.A.

I.N. 250354