Graduate Record Examinations®

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

2001-2002

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

Visit GRE Online at www.gre.org

ETS® Published for the Graduate Record Examinations Board by Educational Testing Service

Overview

The concern for fairness pervades all aspects of testing, including the (1) development of the tests, (2) standardization of testing conditions, and (3) 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 GRE tests and test scores are fair for all test takers, regardless of group membership. The purpose of this publication is to discuss (1) the history of the GRE General Test; (2) the development of the GRE General Test; (3) observed GRE General Test mean score differences; (4) the procedures ETS follows to ensure that its tests are fair to all individuals regardless of group membership; (5) GRE research on validity of the GRE General Test with respect to different sex, age, racial, and ethnic groups; and (6) specific score-use guidelines that are especially important in the context of test fairness.

History of the GRE General Test

The Graduate Record Examinations are an outgrowth of a project funded by The Carnegie Foundation for the Advancement of Teaching in the early 1930s to study the outcomes of college education. However, widespread use of the GRE General Test did not begin until after World War II, when a much larger and more diverse student body began to pursue graduate degrees. The test was used by institutions as a common, objective measure to evaluate the credentials of applicants from differing, and often not widely known, undergraduate programs. To provide a better basis for evaluating students, test results were used to supplement other evidence of students' qualifications. Therefore, the test helped to promote greater fairness and equity than was likely through existing admissions procedures. Today use of the test continues to enhance equity, fairness, and access to graduate school.

Development of the GRE General Test

The GRE Board consists of graduate deans and other members of the graduate education community. The Board defines the content of the GRE General Test as a measure of knowledge and skills that members of the graduate community have identified as important for graduate study — for example, the ability to read with comprehension, to perform basic mathematical operations, to interpret data, to think logically, and to infer relationships. The Technical Advisory Committee for the GRE General Test, which consists of faculty members and deans from various graduate institutions, works with GRE staff to make recommendations to the GRE Board concerning modifications of the test content. Test specialists at ETS are responsible for determining the content of specific test questions and for assembling the General Test.

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 1999-00 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 and analytical measures, and lower on the verbal measure.

Table A.2 presents GRE General Test score information by ethnic group status and sex for U.S. citizens for the 1999-00 testing year. 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.

Tables A.3, A.4, and A.5 present mean GRE General Test scores by broad intended graduate major field and sex for the 1999-00 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 1999-00 testing year. In these tables, mean scores vary considerably by graduate major field. Within each major field and measure, performance differences among gender and racial/ethnic groups

This publication can be downloaded from the GRE Web site at www.gre.org/codelst.html.

are noted. The magnitudes of these differences vary by major field and measure.

Differences in average scores of certain groups do not necessarily mean that the test is biased or favors one group over another. Group differences in performance can result, in part, from group differences in early education and undergraduate course-taking patterns. Male and female students often differ in their interests as well as in their educational experiences. It would be surprising if such differences were not reflected in performance on broad-based educational tests.

Group differences may reflect the unequal knowledge and skills resulting from different educational, economic, and social systems in which everyone does not receive equal opportunity. It is important that tests identify this inequality; such test information can help educators identify and correct deficiencies that can impede success in advanced studies. Further instruction could decrease or eliminate the differences.

The fact that meaningful educational differences exist, however, does not relieve test developers of the obligation to ensure, to the extent possible, that test questions are fair to all test takers. It is crucial that everything possible be done to ensure that tests are fair to everyone.

Steps That ETS Takes to Ensure Fairness

ETS has designed several procedures intended to build fairness into its tests: involving external faculty members in the design and oversight of the tests, the fairness review process, and the differential item functioning (DIF) analysis. The purpose of involving faculty members in the design and oversight of the tests is to make sure that the perspectives of a diverse group of people are considered in planning and ongoing operational activities. The purpose of the fairness review process is to ensure that tests reflect the multicultural nature of society, and to screen out any material that might be offensive or less accessible 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.

Involving External Faculty Members in the Design and Oversight of the General Test. The GRE Program involves undergraduate and graduate faculty members in the design and oversight of the General Test. The GRE Technical Advisory Committee is made up of men and women from a number of academic disciplines and who represent a variety of kinds of institutions. Members are drawn from a variety of ethnic groups. Geographical diversity is also sought. Drawing on a diverse group of educators, who are not ETS employees, is one way ETS seeks to ensure the fairness of the General Test.

Fairness Review. Every question in an 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 or be removed from the test. The fairness review does not guarantee that women, minority group members, or individuals with disabilities 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 or inaccessible. 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 particular test question. Differential item functioning analysis is a statistical technique used as part of the pretesting process that is designed to identify test questions that are more difficult for members of one group than for members of some other group, controlling for 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 it is 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 similar ability levels*. 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 criterion used to match the two groups on ability. Then the DIF analysis is repeated to see if this improved criterion reveals any additional questions that are particularly difficult for members of the focal group.

When questions are pretested and sample sizes permit, DIF analyses are performed before the questions are selected for the operational test. A question showing a large DIF value will not be included in the test, unless the question is considered essential for the test's content coverage.

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. ETS decided to prohibit for skills tests further

use of questions in those categories, regardless of the DIF performance of particular questions in those categories.

The GRE Program encourages test takers to report concerns about specific test questions directly to the test center administrator or to the GRE Program immediately following the test administration. Subject matter specialists will review these questions and eliminate them from scoring if potential bias is determined. The test specialists will also respond in writing to the examinees. If a response does not resolve an examinee's concern, the examinee may pursue the matter further with ETS.

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 sizes of minority groups are not large enough to be definitive, the available data do not show evidence of bias. The data have shown that the scores generally predict about as well for test takers of one sex as for the other. The data have also shown that the scores generally predict about as well for test takers who communicate better in English as for those who do not communicate better in English.

One exception to this general pattern of results involves older students. When students over age 24 are considered as a separate group, GRE quantitative and analytical scores and undergraduate grades tend to slightly underestimate the students' graduate grades. This underestimation appears particularly true for women over 24, who on the average obtain graduate grades about two-tenths of a grade point higher than those of students 24 or under with the same undergraduate grades and GRE scores.

The tendency of GRE General Test scores to underestimate the graduate grades of older students, particularly women, 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 occurs for other important criteria of graduate school success.

Score Use Guidelines

The guidelines for the use of GRE scores provide 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 are included in the *GRE Guide to the Use of Scores*. Four 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. Nor do the GRE Tests assess some 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 or to accept 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 300 on the verbal measure and 800 on the quantitative measure is very different from an applicant with 800 on verbal and 300 on quantitative. The former applicant might do well in a mathematics program, but the latter probably would not. Similarly, the student with 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 verbal plus quantitative 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 ability. 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 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. GRE Program staff will provide without charge advice on the design of appropriate validation studies. Information about 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, wording, and 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 also addresses this concern by involving a diverse group of faculty members in the design and oversight of its tests. To address the second type of concern, ETS conducts and publishes statistical studies of the relationships between GRE scores and grades for various groups of test takers and by publishing guidelines for appropriate use of the scores. Taken as a whole, the guidelines for the use of GRE scores promote 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, both the GRE Program and GRE score users have a responsibility for ensuring test use that is not discriminatory on the basis of sex, race, or ethnicity.

Additional questions about policies related to the use and interpretation of GRE scores should be directed to the GRE Program, Educational Testing Service, Princeton, NJ, 08541.

Appendix A

Table A.1

GRE General Test Score Information by Citizenship Status and Sex: 1999-00

			GRE General Test Score Information					
	Examinees		Verbal		Quantitative		Analytical	
Group	Number	Percent*	Mean	SD	Mean	SD	Mean	SD
U.S. Citizens	239,071	71	478	107	535	133	554	137
Men	82,688	25	496	109	582	134	568	141
Women	156,383	46	468	104	509	125	546	135
Non-U.S. Citizens	100,260	29	434	131	682	125	581	148
Men	61,441	11	434	131	704	110	587	148
Women	38,819	18	433	130	649	138	573	148
Total	339,331	100	465	116	578	147	562	141
Men	144,129	42	470	123	634	138	576	144
Women	195,202	58	461	111	537	140	551	138

Note: A total of 363,932 examinees took the GRE General Test in 1999-00, and 93 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: 1999-00
(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	1,466	1	453	101	485	131	513	137
Men	511	<1	471	101	524	138	526	140
Women	955	<1	444	100	465	122	505	134
Asian/Pacific	10,821	5	480	117	610	129	571	140
Men	4,189	2	485	120	653	120	583	143
Women	6,632	3	476	115	584	127	563	137
Black/African	21,208	9	392	91	419	123	427	118
Men	5,588	2	399	97	451	136	430	124
Women	15,620	7	389	89	407	115	426	115
Mexican American	5,352	2	428	98	479	129	488	131
Men	1,912	1	441	102	516	134	494	136
Women	3,440	1	421	96	459	121	484	128
Puerto Rican	2,451	1	402	105	475	131	465	131
Men	878	<1	406	108	512	134	466	136
Women	1,573	<1	400	103	454	125	465	128
Other Hispanic	5,548	2	437	103	487	132	491	139
Men	1,776	1	457	107	539	136	505	144
Women	3,772	1	428	100	463	124	485	136
White	183,419	77	490	102	547	126	572	129
Men	64,088	27	508	103	593	126	584	133
Women	119,331	50	481	100	522	118	566	127
Other	6,640	3	500	117	547	137	559	141
Men	2,760	1	518	117	587	138	573	145
Women	3,880	2	488	115	519	130	549	138
Total	236,905	100	477	107	534	133	554	137
Men	81,702	34	495	106	582	134	568	141
Women	155,203	66	468	104	509	125	546	134

Note: A total of 239,071 U.S. citizens took the GRE General Test in 1999-00, 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: 1999-00

Graduate Major		Men	Women	No Response	Total
Business	N	4,420	5,342	82	9,844
	Mean	451	441	429	446
	SD	111	110	125	111
Education	N	5,264	18,913	141	24,318
	Mean	426	425	427	425
	SD	94	90	99	91
Engineering	N	31,323	7,857	424	39,604
	Mean	454	458	434	454
	SD	125	123	129	125
Humanities and Arts	N	15,360	25,256	246	40,862
	Mean	537	518	491	525
	SD	118	119	125	119
Life Science	N	20,326	47,356	348	68,030
	Mean	465	454	444	457
	SD	108	99	109	102
Physical Science	N	20,736	11,835	224	32,795
	Mean	476	469	425	473
	SD	129	122	122	126
Social Science	N	19,223	42,358	315	61,896
	Mean	486	467	457	473
	SD	115	106	122	109
Other Fields	N	11,185	22,258	254	33,697
	Mean	460	448	431	452
	SD	115	106	117	109
No Response	N	21,439	17,487	13,960	52,886
	Mean	437	436	436	436
	SD	130	123	126	127
Total	N	149,276	198,662	15,994	363,932
	Mean	468	461	437	463
	SD	123	111	126	117

Table A.4

Mean GRE General Test Quantitative Scores by Intended Graduate Major Field and Sex: 1999-00

Graduate Major		Men	Women	No Response	Total
Business	N	4,420	5,342	82	9,844
	Mean	581	528	551	552
	SD	137	148	177	146
Education	N	5,264	18,913	141	24,318
	Mean	502	463	443	471
	SD	126	116	123	119
Engineering	N	31,323	7,857	424	39,604
	Mean	720	704	714	717
	SD	83	89	88	85
Humanities and Arts	N	15,360	25,256	246	40,862
	Mean	558	524	505	537
	SD	131	127	145	130
Life Science	N	20,326	47,356	348	68,030
	Mean	605	540	569	560
	SD	123	126	141	129
Physical Science	N	20,736	11,835	224	32,795
	Mean	707	664	679	691
	SD	99	116	120	108
Social Science	N	19,223	42,358	315	61,896
	Mean	572	517	514	534
	SD	139	129	154	135
Other Fields	N	11,185	22,258	254	33,697
	Mean	559	497	508	518
	SD	139	133	153	138
No Response	N	21,439	17,487	13,960	52,886
	Mean	664	573	643	628
	SD	141	169	140	156
Total	N	149,276	198,662	15,994	363,932
	Mean	635	537	635	582
	SD	139	141	145	148

Table A.5

Mean GRE General Test Analytical Scores by Intended Graduate Major Field and Sex: 1999-00

Graduate Major		Men	Women	No Response	Total
Business	N	4,420	5,342	82	9,844
	Mean	532	523	502	527
	SD	150	148	171	149
Education	N	5,264	18,913	141	24,318
	Mean	490	491	449	490
	SD	133	130	126	130
Engineering	N	31,323	7,857	424	39,604
	Mean	607	625	578	610
	SD	140	132	149	139
Humanities and Arts	N	15,360	25,256	246	40,862
	Mean	568	564	510	565
	SD	142	137	144	139
Life Science	N	20,326	47,356	348	68,030
	Mean	565	555	529	558
	SD	141	134	139	136
Physical Science	N	20,736	11,835	224	32,795
	Mean	617	611	575	615
	SD	141	135	146	139
Social Science	N	19,223	42,358	315	61,896
	Mean	556	552	510	553
	SD	145	136	144	139
Other Fields	N	11,185	22,258	254	33,697
	Mean	536	528	487	530
	SD	142	137	132	139
No Response	N	21,439	17,487	13,960	52,886
	Mean	562	533	555	550
	SD	156	158	146	155
Total	N	149,276	198,662	15,994	363,932
	Mean	574	550	551	560
	SD	147	141	147	144

Table A.6

Mean GRE General Test Verbal Scores by Intended Graduate Major Field and Ethnic Group: 1999-00
(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	51	263	1,104	161	112	176	5,070	123	33	7,093
	Mean	454	452	375	404	395	415	474	441	452	453
	SD	79	117	85	87	111	94	95	111	109	102
Education	N	140	316	2,225	586	149	586	18,288	275	68	22,633
	Mean	427	408	361	380	380	388	440	424	419	428
	SD	92	87	72	80	82	80	87	94	112	89
Engineering	N	47	1,496	814	266	291	305	9,288	359	96	12,962
	Mean	502	482	433	440	385	444	516	511	543	501
	SD	85	122	89	88	89	101	92	116	107	101
Humanities and Arts	N	177	1,073	1,891	757	230	691	27,863	1,287	293	34,262
	Mean	521	536	441	467	462	490	549	547	561	539
	SD	103	118	104	114	128	118	103	113	108	109
Life Science	N	331	2,862	4,064	1,017	525	1,111	44,409	1,189	276	55,784
	Mean	438	473	392	428	388	434	470	479	499	463
	SD	92	108	84	92	94	92	91	109	109	95
Physical Science	N	83	1,106	1,177	259	200	269	12,836	518	178	16,626
	Mean	473	486	404	456	383	478	522	529	566	509
	SD	116	128	94	107	104	106	100	112	114	108
Social Science	N	333	2,080	5,486	1,410	518	1,391	37,860	1,616	279	50,973
	Mean	454	492	396	438	421	441	493	499	530	479
	SD	95	107	93	93	103	99	99	111	120	104
Other Fields	N	216	1,065	2,940	604	274	678	20,501	805	140	27,223
	Mean	431	466	381	415	406	427	473	493	508	460
	SD	101	110	89	93	106	101	99	113	113	104
No Response	N	88	560	1,507	292	152	341	7,304	468	803	11,515
	Mean	427	425	358	387	376	398	470	464	479	448
	SD	103	127	82	92	106	107	103	128	121	112
Total	N	1,466	10,821	21,208	5,352	2,451	5,548	183,419	6,640	2,166	239,071
	Mean	453	480	392	428	402	437	490	500	509	478
	SD	101	117	91	98	105	103	102	117	121	107

Table A.7

Mean GRE General Test Quantitative Scores by Intended Graduate Major Field and Ethnic Group: 1999-00
(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	51	263	1,104	161	112	176	5,070	123	33	7,093
	Mean	479	587	399	443	452	469	539	489	520	512
	SD	109	130	106	107	125	129	113	140	154	126
Education	N	140	316	2,225	586	149	586	18,288	275	68	22,633
	Mean	445	494	375	414	406	417	481	453	447	466
	SD	117	122	97	102	100	101	109	116	126	113
Engineering	N	47	1,496	814	266	291	305	9,288	359	96	12,962
	Mean	584	714	598	643	612	648	702	702	711	692
	SD	102	81	112	98	93	93	80	86	85	89
Humanities and Arts	N	177	1,073	1,891	757	230	691	27,863	1,287	293	34,262
	Mean	492	577	420	470	452	486	538	531	557	530
	SD	127	119	115	117	135	125	118	128	128	123
Life Science	N	331	2,862	4,064	1,017	525	1,111	44,409	1,189	276	55,784
	Mean	493	606	437	496	481	511	550	559	579	542
	SD	118	119	114	123	113	122	113	122	134	119
Physical Science	N	83	1,106	1,177	259	200	269	12,836	518	178	16,626
	Mean	594	690	531	623	547	618	670	667	706	659
	SD	121	97	125	112	111	130	101	113	90	111
Social Science	N	333	2,080	5,486	1,410	518	1,391	37,860	1,616	279	50,973
	Mean	471	577	405	473	436	472	531	532	572	515
	SD	123	121	113	119	122	119	117	127	137	125
Other Fields	N	216	1,065	2,940	604	274	678	20,501	805	140	27,223
	Mean	440	561	388	448	437	452	509	512	533	494
	SD	112	128	110	118	124	117	117	127	150	124
No Response	N	88	560	1,507	292	152	341	7,304	468	803	11,515
	Mean	468	557	370	420	415	432	508	495	532	487
	SD	152	146	112	114	123	139	127	147	142	139
Total	N	1,466	10,821	21,208	5,352	2,451	5,548	183,419	6,640	2,166	239,071
	Mean	485	610	419	479	475	487	547	547	566	535
	SD	131	129	123	129	131	132	126	137	146	133

Table A.8

Mean GRE General Test Analytical Scores by Intended Graduate Major Field and Ethnic Group: 1999-00
(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	51	263	1,104	161	112	176	5,070	123	33	7,093
	Mean	482	535	399	429	452	459	547	482	518	515
	SD	120	145	109	117	136	139	129	148	151	139
Education	N	140	316	2,225	586	149	586	18,288	275	68	22,633
	Mean	479	487	388	426	418	432	512	477	463	494
	SD	126	131	99	106	108	114	121	131	141	125
Engineering	N	47	1,496	814	266	291	305	9,288	359	96	12,962
	Mean	611	615	512	559	506	541	644	629	633	624
	SD	129	136	129	129	130	137	117	128	148	128
Humanities and Arts	N	177	1,073	1,891	757	230	691	27,863	1,287	293	34,262
	Mean	547	575	447	500	483	521	586	574	600	574
	SD	135	133	121	131	140	140	127	135	140	132
Life Science	N	331	2,862	4,064	1,017	525	1,111	44,409	1,189	276	55,784
	Mean	519	571	440	508	464	508	573	559	570	559
	SD	133	137	115	129	124	134	124	137	146	130
Physical Science	N	83	1,106	1,177	259	200	269	12,836	518	178	16,626
	Mean	569	608	479	550	480	568	637	624	666	619
	SD	129	135	125	129	130	146	120	135	115	130
Social Science	N	333	2,080	5,486	1,410	518	1,391	37,860	1,616	279	50,973
	Mean	520	570	432	495	471	495	574	558	598	552
	SD	131	133	117	129	128	137	127	136	141	135
Other Fields	N	216	1,065	2,940	604	274	678	20,501	805	140	27,223
	Mean	471	544	411	473	451	470	551	540	554	530
	SD	135	134	111	127	138	132	127	138	144	135
No Response	N	88	560	1,507	292	152	341	7,304	468	803	11,515
	Mean	458	491	376	417	408	422	525	497	531	495
	SD	145	152	103	117	118	132	133	146	147	142
Total	N	1,466	10,821	21,208	5,352	2,451	5,548	183,419	6,640	2,166	239,071
	Mean	513	571	427	488	465	491	572	559	569	554
	SD	137	140	118	131	131	139	129	141	149	137

Appendix B

The ETS Fairness Review Process

Reviewers

Reviews of ETS publications are conducted by ETS staff members who are specifically trained in fairness issues at one-day workshops, which are supplemented with periodic refresher courses and the advice of experienced mentors. All staff who write, review, or produce test assessments and publications, or who conduct research, receive this training. In addition, non-ETS staff members who review test questions and test forms are trained in fairness issues.

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 arbitration process. A preliminary review is an excellent means of identifying potential problems early, when modification can be made easily. The mandatory review occurs when the document or assessment is in final form. If a writer and the fairness reviewer disagree about the material, and the disagreement cannot be resolved to mutual satisfaction, an arbitration process occurs in which a panel of staff members who are not involved with the material makes a final determination about what is acceptable.

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 might reinforce 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*. Whereas 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 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. The reason for this exclusion is that controversial material may distract some examinees, thereby reducing their performance on the test.
5. *Contextual considerations*. Sometimes the use of potentially sensitive material is unavoidable. There are four main areas in which this may occur:
   - *Historical domain:* 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 African American 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 or slang.

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, NJ: Erlbaum, 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.

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 ability as estimated by performance on tests and subtests. These ability estimates have been shown to be reliable and valid, and they are obtained under standardized conditions for all examinees. Even though people with the same performance level are not identical, they are likely to be reasonably well matched in terms of the knowledge and skill measured by the test.

The procedure looks within each cluster of people at a single ability 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 ability level and 16 of them answer correctly, the odds are 16/4 or 4 to 1 that a woman at that ability level will answer correctly. If 12 out of 18 men answer the questions correctly, the odds are 12/6 or 2 to 1 that a man at that ability level will answer the question correctly.

After each ability level has been analyzed, there is a calculation of the ratio of the two odds to obtain an indication of the relative advantage of one group over the other within the ability 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 ability level. The "odds ratios" are then averaged across all of the ability 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] 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.

Standardized P-Difference

The other DIF statistic in routine use at ETS is called the Standardized P-Difference. To compute this statistic, all the examinees in each of the two groups being compared are classified according to their ability levels. At each ability level, 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.) is computed. The difference between these two proportions at each ability level is computed. Then the data for all the ability levels are combined in the following way: (1) the difference between groups at each ability level is multiplied by the percentage of the focal group scoring at that level; (2) these weighted differences are combined 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 the difference between groups is computed separately at each ability level, using all available focal group and reference group examinees. Then the differences over all the ability levels are averaged 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 ability levels with the greatest concentration of focal group members.

[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.

50% RECYCLED PAPER
10% Post Consumer Waste

54053-06387 • U81M15 • Printed in the U.S.A.

I.N. 992298