

**Sex,
Race,
Ethnicity,
and Performance
on the
GRE®
General Test
2003-2004**

IMPORTANT

This publication provides GRE General Test score data from the 2001-2002 testing year. Therefore, this publication does not include score data on the new analytical writing section that became part of the General Test beginning in October 2002.

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

Visit GRE online at www.gre.org



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 from the 2001-02 testing year are presented. Table A.1 presents GRE General Test score information by citizenship status and sex for the 2001-02 testing year. These data show that men generally have higher mean scores. Non-U.S. women have a higher mean score on the verbal 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 2001-02 testing year. In Table A.2, test takers who identified themselves as White, Asian/Pacific, and Other have higher mean scores on the verbal, quantitative, and analytical measures than do the other ethnic groups. In addition, within each ethnic group, mean scores are higher for men than women on the verbal, quantitative, and analytical measures.

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

Copyright © 2003 by Educational Testing Service. All rights reserved.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logos, GRADUATE RECORD EXAMINATIONS, and GRE are registered trademarks of Educational Testing Service.

Tables A.3, A.4, and A.5 present mean GRE General Test scores by broad intended graduate major field and sex for the 2001-02 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 2001-02 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 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 different 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 writing 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: 2001-02

Group	GRE General Test Score Information							
	Examinees		Verbal		Quantitative		Analytical	
	Number	Percent*	Mean	SD	Mean	SD	Mean	SD
U.S. Citizens	282,615	66	481	108	545	137	558	135
Men	100,055	23	500	110	594	136	572	138
Women	182,560	43	471	106	518	130	550	133
Non-U.S. Citizens	145,931	34	457	146	700	120	597	143
Men	88,126	21	453	145	715	107	598	142
Women	57,805	13	464	147	677	134	595	144
Total	428,546	100	473	123	597	151	571	139
Men	188,181	44	478	130	650	138	584	140
Women	240,365	56	469	117	556	148	561	137

Note: A total of 462,528 examinees took the GRE General Test in 2001-02, 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: 2001-02
(U.S. Citizens Only)*

Group ¹	GRE General Test Score Information							
	Examinees		Verbal		Quantitative		Analytical	
	Number	Percent ²	Mean	SD	Mean	SD	Mean	SD
American Indian	1,611	1	453	101	496	133	512	132
Men	561	<1	473	105	540	137	517	136
Women	1050	<1	443	97	472	125	509	130
Asian/Pacific	14,379	5	487	121	628	131	580	138
Men	5,849	2	490	124	671	120	592	142
Women	8,530	3	484	119	599	131	572	135
Black/African	25,007	9	394	92	425	128	435	118
Men	6,582	2	401	97	458	140	436	123
Women	18,425	7	391	90	413	122	434	116
Mexican American	6,661	2	428	99	483	133	488	129
Men	2,311	1	442	100	530	138	501	133
Women	4,350	1	420	98	459	123	481	126
Puerto Rican	2,939	1	408	105	478	132	478	127
Men	1,010	<1	419	112	516	138	479	133
Women	1,929	<1	403	100	458	125	477	124
Other Hispanic	6,820	2	441	104	499	137	500	134
Men	2,294	1	459	110	550	138	514	141
Women	4,526	1	431	100	472	129	492	130
White	212,912	76	494	103	556	129	576	127
Men	76,178	27	512	103	603	128	587	131
Women	136,734	49	484	101	531	122	569	125
Other	9,527	3	505	116	561	140	565	137
Men	3,978	1	524	115	605	135	577	138
Women	5,549	2	492	115	530	135	555	135
Total	279,856	100	481	108	544	137	558	135
Men	98,763	35	500	110	593	136	571	138
Women	181,093	65	471	105	517	130	550	133

Note: A total of 282,615 U.S. citizens took the GRE General Test in 2001-02, and 99 percent responded to questions in this table.

¹Ethnic groups are defined as follows: American Indian: American Indian or Alaskan Native; Asian/Pacific: Asian, Asian American, or Pacific Islander; Black/African: Black or African American; Mexican American: Mexican, Mexican American, or Chicano; Puerto Rican-same; Other Hispanic: Other Hispanic or Latin American; White-White (non-Hispanic); Other-same.

²Percentages in this table are based on the column total.

Table A.3*Mean GRE General Test Verbal Scores by Intended Graduate Major Field and Sex: 2001-02*

Graduate Major		Men	Women	No Response	Total
Business	{ N	6,236	7,976	65	14,277
	{ Mean	459	451	445	454
	{ SD	118	116	133	117
Education	{ N	5,857	20,553	167	26,577
	{ Mean	429	426	438	427
	{ SD	96	92	101	93
Engineering	{ N	41,586	11,751	400	53,737
	{ Mean	464	478	445	467
	{ SD	133	133	129	133
Humanities and Arts	{ N	18,579	31,656	273	50,508
	{ Mean	544	525	504	532
	{ SD	116	118	130	118
Life Science	{ N	23,195	51,402	403	75,000
	{ Mean	473	459	435	463
	{ SD	117	105	116	109
Physical Science	{ N	29,902	16,569	243	46,714
	{ Mean	489	482	447	486
	{ SD	136	131	147	135
Social Science	{ N	23,047	50,709	308	74,064
	{ Mean	494	473	459	479
	{ SD	118	109	122	112
Other Fields	{ N	13,869	28,402	221	42,492
	{ Mean	470	455	456	460
	{ SD	118	110	124	113
No Response	{ N	33,379	26,213	19,567	79,159
	{ Mean	443	454	438	446
	{ SD	140	138	129	137
Total	{ N	195,650	245,231	21,647	462,528
	{ Mean	476	469	440	470
	{ SD	130	117	129	124

Table A.4*Mean GRE General Test Quantitative Scores by Intended Graduate Major Field and Sex: 2001-02*

Graduate Major		Men	Women	No Response	Total
Business	{ N	6,236	7,976	65	14,277
	{ Mean	595	549	535	569
	{ SD	141	153	152	150
Education	{ N	5,857	20,553	167	26,577
	{ Mean	508	471	456	479
	{ SD	132	124	128	127
Engineering	{ N	41,586	11,751	400	53,737
	{ Mean	727	716	724	725
	{ SD	84	88	81	85
Humanities and Arts	{ N	18,579	31,656	273	50,508
	{ Mean	566	537	509	548
	{ SD	136	135	154	136
Life Science	{ N	23,195	51,402	403	75,000
	{ Mean	623	556	562	577
	{ SD	129	134	144	136
Physical Science	{ N	29,902	16,569	243	46,714
	{ Mean	715	678	682	702
	{ SD	99	116	110	107
Social Science	{ N	23,047	50,709	308	74,064
	{ Mean	587	529	504	547
	{ SD	143	136	155	141
Other Fields	{ N	13,869	28,402	221	42,492
	{ Mean	570	507	498	527
	{ SD	142	140	166	144
No Response	{ N	33,379	26,213	19,567	79,159
	{ Mean	679	609	653	649
	{ SD	133	169	138	150
Total	{ N	195,650	245,231	21,647	462,528
	{ Mean	651	557	646	601
	{ SD	138	149	142	151

Table A.5*Mean GRE General Test Analytical Scores by Intended Graduate Major Field and Sex: 2001-02*

Graduate Major		Men	Women	No Response	Total
Business	{ N	6,236	7,976	65	14,277
	{ Mean	542	537	485	539
	{ SD	148	147	159	148
Education	{ N	5,857	20,553	167	26,577
	{ Mean	491	497	457	496
	{ SD	133	129	137	130
Engineering	{ N	41,586	11,751	400	53,737
	{ Mean	613	632	602	617
	{ SD	133	126	121	132
Humanities and Arts	{ N	18,579	31,656	273	50,508
	{ Mean	573	570	509	571
	{ SD	141	136	148	138
Life Science	{ N	23,195	51,402	403	75,000
	{ Mean	571	563	515	565
	{ SD	140	133	140	136
Physical Science	{ N	29,902	16,569	243	46,714
	{ Mean	622	617	573	620
	{ SD	136	133	138	135
Social Science	{ N	23,047	50,709	308	74,064
	{ Mean	564	558	501	559
	{ SD	145	136	148	139
Other Fields	{ N	13,869	28,402	221	42,492
	{ Mean	544	534	491	537
	{ SD	141	136	145	138
No Response	{ N	33,379	26,213	19,567	79,159
	{ Mean	573	557	566	566
	{ SD	146	153	137	147
Total	{ N	195,650	245,231	21,647	462,528
	{ Mean	582	559	562	569
	{ SD	143	140	138	142

Table A.6

*Mean GRE General Test Verbal Scores by Intended Graduate Major Field and Ethnic Group: 2001-02
(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	67	416	1,543	237	112	282	7,037	231	43	9,968
	{ Mean	451	460	377	414	385	422	476	462	500	456
	{ SD	81	121	81	91	96	93	93	102	95	100
Education	{ N	138	346	2,380	658	178	604	19,704	369	59	24,436
	{ Mean	414	418	366	378	374	383	438	435	411	427
	{ SD	104	89	75	81	84	80	87	104	86	89
Engineering	{ N	83	2,362	1,095	373	278	467	11,960	555	129	17,302
	{ Mean	462	490	430	439	405	454	514	512	542	500
	{ SD	96	120	94	89	94	100	92	116	125	101
Humanities and Arts	{ N	199	1,478	2,131	965	304	859	34,145	1,841	367	42,289
	{ Mean	505	550	441	469	478	490	551	555	575	542
	{ SD	99	121	105	111	125	112	102	104	105	107
Life Science	{ N	338	3,015	4,367	1,026	596	1,136	45,468	1,505	250	57,701
	{ Mean	442	474	394	424	392	430	472	480	508	464
	{ SD	89	112	82	91	94	95	92	107	119	96
Physical Science	{ N	95	1,835	1,600	351	262	396	16,341	806	221	21,907
	{ Mean	488	486	408	455	387	465	526	528	566	511
	{ SD	110	132	92	105	103	112	101	121	97	110
Social Science	{ N	390	2,764	6,251	1,770	689	1,779	44,115	2,340	354	60,452
	{ Mean	456	501	398	436	423	450	496	501	537	483
	{ SD	95	114	94	97	99	101	100	111	118	106
Other Fields	{ N	202	1,394	3,693	885	327	876	25,570	1,235	173	34,355
	{ Mean	443	473	383	413	411	433	477	498	504	464
	{ SD	103	115	91	91	105	98	100	114	117	105
No Response	{ N	99	769	1,947	396	193	421	8,572	645	1,163	14,205
	{ Mean	411	427	363	387	373	400	473	471	485	450
	{ SD	110	122	86	91	100	103	104	128	125	114
Total	{ N	1,611	14,379	25,007	6,661	2,939	6,820	212,912	9,527	2,759	282,615
	{ Mean	453	487	394	428	408	441	494	505	515	481
	{ SD	101	121	92	99	105	104	103	116	123	108

Table A.7

*Mean GRE General Test Quantitative Scores by Intended Graduate Major Field and Ethnic Group: 2001-02
(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	67	416	1,543	237	112	282	7,037	231	43	9,968
	{ Mean	517	595	402	469	428	495	549	535	583	523
	{ SD	115	131	117	115	119	123	116	140	122	130
Education	{ N	138	346	2,380	658	178	604	19,704	369	59	24,436
	{ Mean	439	507	386	416	405	425	484	462	447	471
	{ SD	125	119	104	109	115	108	113	122	126	117
Engineering	{ N	83	2,362	1,095	373	278	467	11,960	555	129	17,302
	{ Mean	650	725	603	657	616	660	707	704	727	699
	{ SD	107	78	114	99	102	92	80	89	70	89
Humanities and Arts	{ N	199	1,478	2,131	965	304	859	34,145	1,841	367	42,289
	{ Mean	482	590	420	468	468	483	546	543	567	537
	{ SD	119	125	119	120	139	130	121	128	125	126
Life Science	{ N	338	3,015	4,367	1,026	596	1,136	45,468	1,505	250	57,701
	{ Mean	499	616	439	506	486	511	557	572	596	549
	{ SD	121	122	118	124	117	125	115	129	138	122
Physical Science	{ N	95	1,835	1,600	351	262	396	16,341	806	221	21,907
	{ Mean	620	699	539	620	547	632	678	681	714	666
	{ SD	116	102	130	118	121	117	102	113	96	113
Social Science	{ N	390	2,764	6,251	1,770	689	1,779	44,115	2,340	354	60,452
	{ Mean	475	597	412	471	454	482	542	543	581	526
	{ SD	126	127	121	122	116	125	121	132	143	130
Other Fields	{ N	202	1,394	3,693	885	327	876	25,570	1,235	173	34,355
	{ Mean	464	572	389	454	442	468	517	527	521	502
	{ SD	124	129	111	117	127	124	121	130	144	128
No Response	{ N	99	769	1,947	396	193	421	8,572	645	1,163	14,205
	{ Mean	475	572	377	419	427	434	516	515	554	496
	{ SD	146	154	115	121	125	136	131	154	142	143
Total	{ N	1,611	14,379	25,007	6,661	2,939	6,820	212,912	9,527	2,759	282,615
	{ Mean	496	628	425	483	478	499	556	561	580	545
	{ SD	133	131	128	133	132	137	129	140	146	137

Table A.8

*Mean GRE General Test Analytical Scores by Intended Graduate Major Field and Ethnic Group: 2001-02
(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	67	416	1,543	237	112	282	7,037	231	43	9,968
	{ Mean	527	547	410	461	443	481	552	529	590	524
	{ SD	123	142	110	121	121	129	125	140	120	135
Education	{ N	138	346	2,380	658	178	604	19,704	369	59	24,436
	{ Mean	449	486	402	420	426	437	515	492	460	497
	{ SD	127	122	101	107	114	116	118	132	124	123
Engineering	{ N	83	2,362	1,095	373	278	467	11,960	555	129	17,302
	{ Mean	577	628	519	556	514	559	641	628	652	625
	{ SD	136	131	128	131	128	132	117	135	123	127
Humanities and Arts	{ N	199	1,478	2,131	965	304	859	34,145	1,841	367	42,289
	{ Mean	517	586	451	494	503	513	589	579	599	577
	{ SD	131	134	119	131	139	134	125	126	126	130
Life Science	{ N	338	3,015	4,367	1,026	596	1,136	45,468	1,505	250	57,701
	{ Mean	518	571	446	507	479	510	576	560	592	561
	{ SD	124	134	114	124	119	134	122	132	133	128
Physical Science	{ N	95	1,835	1,600	351	262	396	16,341	806	221	21,907
	{ Mean	574	607	480	543	486	550	636	621	669	617
	{ SD	127	139	122	137	127	140	119	135	126	131
Social Science	{ N	390	2,764	6,251	1,770	689	1,779	44,115	2,340	354	60,452
	{ Mean	509	579	439	498	486	505	578	560	600	557
	{ SD	127	132	119	126	128	130	125	135	141	133
Other Fields	{ N	202	1,394	3,693	885	327	876	25,570	1,235	173	34,355
	{ Mean	505	551	418	475	470	488	554	551	545	535
	{ SD	133	131	110	121	123	128	125	131	146	131
No Response	{ N	99	769	1,947	396	193	421	8,572	645	1,163	14,205
	{ Mean	463	508	381	426	426	435	531	505	539	501
	{ SD	139	150	102	120	120	125	131	149	139	141
Total	{ N	1,611	14,379	25,007	6,661	2,939	6,820	212,912	9,527	2,759	282,615
	{ Mean	512	580	435	488	478	500	576	565	575	558
	{ SD	132	138	118	129	127	134	127	137	143	135

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.

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



54719-06387 • Y63M5 • Printed in the U.S.A.

I.N. 997949