DOCUMENT RESUME

ED 281 863 TM 870 244

AUTHOR Braun, Henry; And Others
TITLE The Predictive Validity of the GRE General Test for Disabled Students. Studies of Admissions Testing and Handicapped People, Report No. 10.
INSTITUTION College Entrance Examination Board, Princeton, N.J.; Educational Testing Service, Princeton, N.J.; Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-86-42
PUB DATE Nov 86
NOTE 44p.; For other reports in this series, see ED 251 485, ED 251 487, ED 268 154, ED 269 418, and ED 275 697.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Bayesian Statistics; *College Entrance Examinations; Correlation; Grade Point Average; Higher Education; *Learning Disabilities; Mathematics Tests; *Physical Disabilities; *Predictive Validity; *Test Theory; Verbal Tests; *Visual Impairments
IDENTIFIERS *Graduate Record Examinations; Residuals (Statistics)

ABSTRACT

From the fall of 1981 through June 1984, more than 850 disabled examinees took special administration of the Graduate Record Examinations (GRE). Grade point averages were obtained on 278 disabled students; 236 had enough complete data to be included in the study. Disabled students earned lower mean GRE scores than their nonhandicapped counterparts but, except for visually impaired students, they earned overall graduate grade point averages very close to those of nonhandicapped students. Differences in the GRE performance of blind students compared to students with other visual disabilities raised questions for further research. The predictive validity of the GRE scores obtained in nonstandard administrations was estimated with empirical Bayes procedures. The differences between actual and predicted grade point averages were more negative for disabled students. The correlations between predicted and actual scores were more modest than results for nonhandicapped students. In addition a distinctive pattern was observed for disabled students: higher predicted five-year averages (FYA) scores were accompanied by increased overprediction. However, considerable caution was recommended in interpreting the results of the study because of the lack of information on nonhandicapped students in the same departments as disabled students. Appendices include a description of the use of Empirical Bayes methods and an annotated bibliography of reports from "Studies of Admissions Testing and Handicapped People." (Author/JAZ)

RR-86-42

THE PREDICTIVE VALIDITY OF THE GRE GENERAL TEST FOR DISABLED STUDENTS

Henry Braun
Marjorie Ragosta
and
Bruce Kaplan

November 1986

Report No. 10
Studies of Admissions Testing and Handicapped People
A Project Sponsored by

College Entrance Examination Board
Educational Testing Service
Graduate Record Examinations Board

Studies of Admissions Testing and Handicapped People

Most admissions testing programs have long made accommodations for handicapped examinees, though practices have varied across programs and limited research has been undertaken to evaluate such test modifications. Regulations under Section 504 of the Rehabilitation Act of 1973 impose new requirements on institutional users, and indirectly on admissions test sponsors and developers, in order to protect the rights of handicapped persons. The Regulations have not been strictly enforced since many have argued that they conflict with present technical capabilities of test developers. In 1982, a Panel appointed by the National Research Council released a detailed report and recommendations calling for research on the validity and comparability of scores for handicapped persons.

Due to a shared concern for these issues, College Board, Educational Testing Service, and Graduate Record Examinations Board initiated a series of studies in June 1983. The primary objectives are:

To develop an improved base of information concerning the testing of handicapped populations.

To evaluate and improve wherever possible the accuracy of assessment for handicapped persons, especially test scaling and predictive validity.

To evaluate and enhance wherever possible the fairness and comparability of tests for handicapped and nonhandicapped examinees.

This is one of a series of reports on the project, which will continue through 1986. Opinions expressed are those of the authors. See Appendix for an annotated bibliography of earlier reports of the series.

ETS Research Report 86-42

The Predictive Validity of the GRE General Test for Disabled Students

Henry Braun

Marjorie Ragosta

Bruce Kaplan

November 1986

Acknowledgements

We wish to thank all of those people whose work contributed to the project, and especially:

Ka-Ling Chan and Tom Jirele for help with the data analysis;

Hazel Klein for her assistance in matching the data;

Shirley Perry for typing and Linda DeLauro for editing the report;

Craig Mills, Donald Rock, and Kenneth Wilson for reviewing the report and Warren Willingham for his many helpful suggestions.

We also wish to express appreciation to the College Board Joint Staff Research and Development Committee and the Graduate Record Examinations Board Research Committee for their financial assistance.

Abstract

From the fall of 1981 through June, 1984, more than 850 disabled examinees took special administrations of the GRE. Through efforts involving more than 400 graduate schools, grade point averages were obtained on 278 of these disabled students, about 236 of whom had data complete enough to be included in the study. Disabled students earned lower mean GRE scores than their nonhandicapped counterparts, but, except for visually impaired students, they earned overall graduate grade point averages very close to those of nonhandicapped students. Differences in the GRE performance of blind students compared to students with other visual disabilities raised questions for further research.

The predictive validity of the GRE scores obtained in nonstandard administrations was estimated with empirical Bayes procedures. The differences between actual and predicted grade point averages were more negative for disabled students, and the correlations between predicted and actual scores more modest than results for nonhandicapped students. In addition a distinctive pattern was observed for disabled students: higher predicted FYA scores were accompanied by increased overprediction. However, considerable caution was recommended in interpreting the results of the study because of the lack of information on nonhandicapped students in the same departments as disabled students.

Introduction

In response to a call by the Panel on Testing of Handicapped People (Sherman & Robinson, 1982) for a program of research, the College Board (CB), Educational Testing Service (ETS), and the Graduate Record Examination Board (GREB) jointly funded a project, "Studies of Admissions Testing and Handicapped People." As part of that research effort, this study presents data on the validity of the Graduate Record Examination (GRE) Aptitude Test as a predictor of college performance for people in four disability classifications: hearing impairment, learning disability, physical handicap or visual impairment. These validity data address the question of whether the GRE predicts the graduate performance of people with disabilities as well as it predicts the performance of graduate students in general.

1. Research Design & Implementation

A major focus of the federal regulations implementing Section 504 of the Rehabilitation Act of 1973 was the predictive validity of admissions tests for disabled test takers. Although validity studies of the GRE have been routinely performed for the general population (Braun & Jones, 1985; Livingston & Turner, 1982; Wilson, 1982) and for some special populations (e.g. minority examinees), no studies have involved specific handicapped groups (Bennett, Ragosta, & Stricker, 1984).

The Panel on Testing of Handicapped People (Sherman & Robinson, 1982) also emphasized validity in its program of research. If it could be shown that all of the modifications made for handicapped people in a given test

produced scores that predicted future performance as well as scores on the regular version did, then an important source of doubt about the appropriateness of the test would disappear. The panel noted the paucity of data and suggested studies of the effects of modifying tests and testing procedures. Recognizing the difficulty of finding enough disabled students within any single institution to provide data for a standard validity study, the panel recommended developing a validation technique which would facilitate the pooling of information across many institutions. With that charge in mind, a research design was developed, incorporating empirical Bayes methodology as the basis of the validation technique. In a recent analysis of the predictive validity of SAT (Scholastic Aptitude Test) scores for disabled students (Braun, Ragosta & Kaplan, 1986), empirical Bayes facilitated the estimation of prediction equations for a relatively large number of schools with relatively small numbers of handicapped people.

In this setting, the implementation of empirical Bayes methods was impeded by the lack of control data from the departments in which the handicapped students in the study were enrolled. Consequently, a rather elaborate strategy was developed in order to obtain estimates of the prediction equations for those departments. The procedure is described in some detail in Appendix A. Because the process does involve substantial extrapolation, the residuals that are derived from these prediction equations must be interpreted with great caution. Moreover, it must be borne in mind that the population of interest, handicapped graduate students that have taken special forms of the GRE, forms a miniscule fraction of the total population of graduate students and that while our sample represents a substantial fraction of that population it is a non-random sample. The combination of factors--the scarcity of the data and the data-analysis strategy requiring substantial extrapolation--gives us less

confidence in the results of this study than we might have desired. Nevertheless, we thought it important to report the results, in part to establish the difficulty of doing validity studies under these conditions, and in part to show that--despite the difficulties--the analyses reveal patterns similar to those discussed in the SAT validity study (Braun, Ragosta & Kaplan, in press).

Design Considerations

The major focus of the current validity study is the scores derived from special test administrations of the GRE. In special administrations, people with disabilities may take a regular or modified version of the GRE--e.g., braille, large type, or cassette versions--under special conditions including a separate location, extra time, a reader, an amanuensis, an interpreter, rest periods, special equipment or other adaptations. During the period from the fall of 1981 through the end of June, 1984, more than 850 examinees took special test administrations of the GRE. At the time of data collection for this study, some proportion of those students could be assumed to be in graduate institutions. Others perhaps never attended graduate school, while still others may have attended and dropped out. Before we could collect data for the validity studies, we needed to locate those people who had been admitted to specific departments of graduate schools after taking special test administrations of the GRE.

A second consideration was the interest in studying a second control group composed of disabled people who had taken regular test administrations

of the GRE and attended the same departments in the same graduate institutions. The immediate task was to identify these handicapped students in departments which had also admitted students with scores from special test administrations.

Given the need for data on two kinds of students, an appropriate data-collection strategy was devised. Existing data files would be searched to determine those graduate schools which had received score reports from special test administrations. For all such graduate institutions, current data files were searched for individuals who had taken regular administrations of the GRE and who had reported having a disability. School by school, lists of the identified handicapped individuals were produced for every graduate school to which students from special test administrations had applied.

Data Collection

Initial contact was made by letter with 432 graduate institutions in June, 1985. The letter requested help for a series of validity studies and included listings of disabled individuals who had sent their GRE scores to the specific graduate school. The listings of students from special administrations typically contained only one or two names. Listings of handicapped students from regular administrations were much longer--often several pages long. To assure a good response, only one or two pages of names were randomly selected and sent to the schools whenever the amount of data requested would have been too burdensome. Graduate institutions were asked to return the forms with validity data for those students who may have attended. Specifically schools were asked to provide the following information:

1. The undergraduate grade-point-average.
2. The first year graduate grade-point-average at the institution.

3. The overall graduate grade-point-average at the institution.

4. The student's graduate department.

A follow-up letter was sent two months after the first contact.

Of the 432 schools which were contacted, 339 responded for a response rate of 78 percent. Of those schools which responded, 188 had no information on students from special GRE administrations. The 151 schools which reported the presence of students from special administrations provided the data for the current study.

With the information obtained from the data collection, a data base was built, composed of information on handicapped students from special test administrations (Specials), and handicapped students from regular test administrations (Regulars). Information on nonhandicapped students in those departments of specific graduate schools containing Specials were not accessible. However, the mean GRE scores of all score senders to the graduate department were available in ETS files and were used in the empirical Bayes model (see Appendix A).

Empirical Bayes Methodology

Estimates of predictive validity are based on obtaining suitable estimates of the regression of some criterion on one or more proposed predictors. In practice, small sample sizes and the effects of self-selection hamper the estimation process. Empirical Bayes methods (Rubin, 1980; Braun, et al., 1983; Braun and Jones, 1985) have been employed with good effect in improving the quality of the validity estimates in a number of different settings, including the predictive validity of the GRE.

With empirical Bayes, a formal mathematical model is developed in which the sets of regression coefficients from different schools are related to one

another. The complexity of the relationship varies from one application to another and the appropriate form may be determined from the data. The most important consequence of this formal model is that it facilitates the "sharing of information" across schools; that is, data from all schools contribute indirectly to the estimation of the regression equation in each school. This sharing of information leads to stable estimates which are superior to the usual least squares estimates based on a single school's data. In fact, the empirical Bayes estimate of a school's prediction equation represents a compromise between the usual least squares estimate and the global estimate based on pooling the data across all schools in the study.

2. Description of the Sample

The data base assembled for this study contained information on 278 students for whom special arrangements were made. Only 273 had GRE scores, and, of these, 37 took the GRE in the standard amount of time, did not have their GRE scores flagged, and did not have disability data available. The remaining Specials were divided among six disability groups: hearing impairment (2), learning disability (29), physical handicap (60), multiple handicaps (15) and visual impairment (40 blind and 90 with other visual impairment). Mean GRE scores, undergraduate grade point averages (UGPA), graduate first-year averages (FYA), and overall graduate grade point averages are presented in Table 2-1.

Insert Table 2-1 about here

The mean GRE scores of these disabled students vary widely. GRE-Verbal means range from a low of 395 for the 15 multiply handicapped students to a high of 518 for the 60 physically handicapped students. The multiply handicapped mean is about one full standard deviation below the mean for physically handicapped people. Blind students had a GRE-Verbal mean of 456 compared to a mean of 506 for students with other (lesser) visual disabilities, a difference of about one-half of a standard deviation. GRE-Quantitative scores ranged from a low of 422 for the 40 blind students to a high of 580 for the two hearing-impaired students. The blind students' GRE-Quantitative mean is again almost half a standared deviation below the mean for students with other visual impairments.

Undergraduate grade point averages range from a low of 2.93 for the 12 multiply handicapped students to a high of 3.27 for the 55 physically handicapped students (and 3.33 for the hearing impaired individual). Multiply handicapped students also had the lowest graduate school means, while the 25 learning disabled students had the highest.

Disabled students in this study were located in 41 of the 99 department codes used to describe major fields (Educational Testing Service, 1983-84). Of the 21 department codes in the Humanities, disabled students were located in only 5, the most popular of which was religion represented by students in 6 different institutions. The Social Sciences were most popular with students in this sample. Of the 27 fields in Social Sciences, disabled students were located in 18, including Clinical Psychology, (9 institutions), Education (5 institutions), Guidance and Counseling (4 institutions), Public Administration (4 institutions), and other Social Sciences (6 institutions). Of the 32 fields in the Biological Sciences, disabled students were enrolled in 10 represented by only one or two institutions each. In the 18 fields of

Physical Sciences, disabled students were located in 8, the most popular of which were Geology (5 institutions) and Chemistry (4 institutions).

About 50 percent of the data on disabled students from regular test administrations came from departments in only 10 schools. The departments represented were primarily in the Social Sciences: Education, Counseling, and Social Welfare. Because of the way these data were collected, they are not representative of the population of disabled students who have taken regular test administrations. Therefore, we have a good deal less confidence in the data from regular administrations than we have in the data from special administrations for disabled students.

The Comprehensive Data Set

The comprehensive validity study required all participants to have UGPA, GRE-Verbal and GRE-Quantitative scores to predict the graduate GPA (either FYA or OA, if FYA was not available). It also required a disability classification for all Specials and the collapsing of categories for visually impaired test takers. Hearing-impaired and multiply handicapped students were dropped because of low numbers. The resultant data base contained three categories of students from special test administrations, one category of handicapped students from regular test administrations, and nonhandicapped controls. The means and standard deviations for these 5 groups on the four variables of the comprehensive validity study are presented in Table 2-2. The data are presented in graphic format in Figure 2-1.

Insert Table 2-2 and Figure 2-1 about here

The test scores of this sample of disabled students taking special test administrations of the GRE are, on the average, lower than the test scores of the nonhandicapped controls. The 19 learning-disabled students in the Comprehensive Data Set had GRE-Verbal and Quantitative scores about one-half of a standard deviation lower than the nonhandicapped controls. The 48 physically handicapped students had GRE-V scores similar to control students but their GRE-Q scores were on the average half a standard deviation lower. The 105 visually impaired students were one-quarter of a standard deviation lower on their GRE-V scores and more than two-thirds of a standard deviation lower on GRE-Q. Despite their lower GRE scores, the disabled Specials had overall graduate grade-point-averages close to that of nonhandicapped students, except for visually impaired individuals whose grades were more than one-third of a standard deviation lower.

3. Data Analysis

In this section we explore the patterns in differential validity across various subgroups of students. In particular we compare distributions of residuals from the estimated prediction equations for nonhandicapped students, handicapped students taking regular administrations, and handicapped students taking special administrations. Three subgroups have been broken out of the latter group: the visually impaired, the learning disabled and the physically handicapped. We investigate three families of prediction equations: one based on employing both GRE scores and UGPA, one GRE scores only and one UGPA only.

As has been made clear in Appendix A, the residual analysis has only been possible after substantial effort had been expended in obtaining an estimate of the prediction equation in each department. This effort was made necessary

by the lack of data on nonhandicapped students in those departments. Since the prediction equations obtained were the result of an extrapolation of a model based on data from other departments, it should be expected that the residuals are at least somewhat more variable than they would have been were relevant control data available. In fact, we show below that the standard deviations of the residuals for the different handicapped groups are typically about fifty percent larger than the standard deviation of the pooled residuals for the nonhandicapped students that were used to estimate the model parameters. This should be compared to our companion study of the SAT (Braun, Ragosta and Kaplan, 1986) in which the corresponding increase is in the range of ten to fifteen percent. Of course, in the latter study relevant control data was available.

It will also be apparent that while the correlations between predicted and actual FYA are not negligible in most cases, the standard deviations of the residuals are comparable in magnitude to the standard deviations of the FYAs. Thus there is considerable variability in the residuals which may mask whatever systematic patterns exist in the data. Some of the excess variability is due to the manner in which the prediction equations were estimated. Some is due undoubtedly to the fact that the groups of handicapped students we study are very heterogeneous and that we have no information on the process by which they matriculated at a given department or the course of study followed. Accordingly, in drawing inferences from these data we will rely on the similarity of the findings to those in the SAT study.

Table 3-1 displays the results for the first family of prediction equations. Recall that nonhandicapped students are culled from 99 departments that participated in the VSS and are in general different from the departments attended by the handicapped students in the study. The prediction equation

employed for each department is the empirical Bayes estimate using the applicant data to provide covariate information. The fit for the nonhandicapped students is very good. The mean residual is close to zero (line 4); moreover, dividing these students into three equal groups according to the level of predicted FYA--low, medium and high--yields only a slight upward trend in the mean residuals (lines 5, 6 and 7). Note also that the correlation between actual and predicted FYA is 0.63, a rather substantial figure. Note, however, that this correlation is obtained from a pooled sample drawn from 99 different departments.

Insert Table 3-1 about here

For the handicapped students, the prediction equations are obtained from the Applicant model as described in Appendix A. The mean residuals are somewhat more negative, -0.06 and -0.09 respectively, and the correlations between actual and predicted FYA are much more modest. Perhaps most important, the pattern in the mean residuals by level of predicted FYA is reminiscent of the patterns observed in the analysis of SAT data (Braun, Ragosta and Kaplan, 1986): increasing predicted FYA is accompanied by increasing overprediction. The trend is quite strong inasmuch as the standard error of one of the subgroup means is approximately 0.06 while the difference in means between the low predicted group and the high predicted group is approximately 0.25 (regular administration) or 0.16 (special administation).

Visually impaired students form the largest subgroup among those taking special administrations and their results mirror those for the group as a

whole. Physically handicapped students form the next largest group and their results are similar to the others with the exception of the trend in the mean residuals (lines 5-7). The results for the learning disabled group are somewhat anomolous with a positive mean residual and a negative correlation between actual and predicted FYA. However, the sample size is so small that it is difficult to lend much credence to these findings.

Table 3-2 presents the results for predictions based on GRE-scores only. For the non-handicapped students the effect is to increase the positive trend in the mean residuals with increasing predicted FYA. For handicapped students, the effect is to make the mean residuals more negative and to further reduce the correlations between actual and predicted FYA. The effect is particularly striking for the learning disabled students.

Insert Table 3-2 about here

Table 3-3 presents the results for predictions based on UGPA only. The pattern of residuals for handicapped students, both from regular and special administrations, indicates substantial overprediction although the correlations between actual and predicted FYA are substantially higher than those in Table 3-2.

Insert Table 3-3 about here

Finally in Table 3-4, we display correlations between residuals and the GRE-V, GRE-Q, UGPA and predicted FYA for various groups of handicapped students. When the residuals are derived from predictions based on test scores and UGPA, they are negatively correlated with the other variables in almost all instances. The negative correlation between the residuals and the predicted FYA is consistent with the pattern in the mean residuals evident in Table 3-1. The correlations are particularly large for the learning disabled students.

Insert Table 3-4 about here

---

4. Discussion

There are two major difficulties in trying to draw conclusions from the preceeding analysis. The first is the extreme caution that must be exercised whenever inferences are drawn from nonrandom samples. As discussed below this is a more serious concern for those disabled students who took a regular administration of the GRE. The second difficulty proceeds from the uncertainty surrounding the estimated prediction equations that were used to generate the residuals for analysis. Because of the lack of relevant control data, the residuals are certainly more variable than they would have been in a more ideal study. Thus, we must be content to observe whatever trends appear and compare them closely with the results obtained from the study of the SAT.

The numbers of disabled students located for the current validity study are small. Although more than 275 disabled students had some validity data, only 216 had complete data for the comprehensive study. More than 60 percent

of that group were students with visual disabilities. In a companion study of the Scholastic Aptitude Test (Braun, Ragosta & Kaplan, in press), visually impaired students were the second largest group of disabled test takers (the largest being the learning-disabled group) but they represented only 20 percent of the population taking special test administrations. Physically handicapped students comprised 27 percent of the GRE sample in the current study, although in the companion SAT study students with physical disabilities were only 9 percent of the sample. Both visually impaired and physically handicapped students take the GRE in larger numbers and relatively larger proportions than students with hearing or learning disabilities. Perhaps the attrition rate in college is higher for students with hearing or learning disabilities.

Despite their low numbers, those disabled students who were Specials included in the comprehensive validity study represent a reasonably large percentage of the total population of students taking special administrations of the GRE. However, the disabled students who took regular administrations of the GRE--the Regulars--are not as representative a group of students as the Specials. Not only do they represent a much smaller proportion of their total population but their composition with regard to disability is unknown. They were an adventitious sample and heavy reliance should not be put on their results.

The difficulty in acquiring sufficient data for a high quality validity study of standardized tests for disabled individuals has been demonstrated again--in this study as well as in the companion study of the SAT. Although the GRE validity study has corroborated some of the findings of the SAT study,

because of small numbers some relevant research could not be done. Testing programs that are much smaller than the GRE might have difficulty in completing a worth-while validity study.

The difference in GRE means between blind and other visually impaired students is worth some follow-up consideration. Why are the GRE scores of blind students significantly lower than the score of students with other visual disabilities? Do blind students have equal access to information and practice materials for the GRE General Test? If so, do blind students have adequate time to take the test in the slower Cassette or Braille versions? Do those students with other visual disabilities have too much time to complete the test? The differences in means between the two groups of students with visual disabilities is too large to ignore. Further research is warranted.

Although the predictive validity findings in this study are less definitive than those in the companion study of the SAT, there is certainly corroborating evidence of overprediction. When GRE scores from special administrations result in relatively high predicted grades in graduate school, those predicted grades are likely to be higher than the grades actually received.

Because of the small numbers of students involved--and the even smaller numbers which would result from categorizing the students by the GRE version used and amount of time taken--no attempt was made to determine whether increased overprediction or higher GRE scores were associated with increased amounts of testing time. Nevertheless, because of the similarity between the results of this study and its companion SAT study with regard to overprediction associated with scores from special test administrations, a similar recommendation is made: that in so far as is possible, the timing of special test administrations become more standardized. It should be borne in

mind, however, that disabled students taking regular administrations of the test were also substantially overpredicted. This is somewhat puzzling although a similar pattern was found in the SAT study. The result does suggest that there may be other factors at work here.

That nonstandard administrations of standardized tests become more standardized with respect to timing was one of the recommendations of the APA-AERA-NCME joint standards (APA, 1985). Although that recommendation seemed inappropriate to us at first--because disabilities are not easily categorized and standardized--the occurrence of overprediction associated with untimed tests has caused us to give more serious thought to the problem of timing. Some further research is necessary to establish suitable timing conditions for students with specific disabilities taking specific versions of the GRE.

Table 2-1

Means and Standard Deviations of Disabled Students
Located for this Study

	GRE-V			GRE-Q			UGPA			FYA			GA		
Special	N	X	SD	N	X	SD	N	X	SD	N	X	SD	N	X	SD
Blind	40	456	110	40	422	124	37	3.13	.46	38	3.23	.72	42	3.24	.66
Other Visual	90	506	106	90	478	131	76	3.11	.56	87	3.34	.50	91	3.37	.52
Physical	60	518	126	60	477	130	55	3.17	.53	61	3.46	.55	63	3.41	.51
LD	29	472	101	29	462	136	24	2.99	.34	25	3.53	.43	26	3.51	.47
Hearing	2	495	95	2	580	120	1	3.33	-	2	3.78	.23	2	3.78	.23
Multiple	15	395	116	15	446	160	12	2.93	.55	15	3.15	.50	15	3.17	.46
Missing*	37	494	122	37	481	165	37	3.14	.48	38	3.44	.42	39	3.46	.39
Total	273	490	118	273	467	139	242	3.11	.51	266	3.38	.54	278	3.38	.52

* These students took the GRE within standard time limits and did not have disability data.

Table 2-2

The Comprehensive Data Set
Means & Standard Deviations for Nonhandicapped and Handicapped Groups

	N	GRE-V		GRE-Q		UGPA		OA	
		X	SD	X	SD	X	SD	X	SD
Nonhandicapped	2025	519	106	543	124	3.26	.44	3.48	.42
Specials									
Learning Disabled	19	469	95	472	142	3.01	.35	3.49	.46
Physically Handicapped	48	514	116	481	129	3.18	.55	3.46	.55
Visually Impaired	105	492	107	457	126	3.09	.53	3.31	.54
Regulars	184	482	122	454	128	3.02	.48	3.40	.50

Table 3-1

Residual Analysis Derived from Predictions
Based on Test Scores and UGPA

	Nonhandicapped	Handicapped				
		Regular	Special			
			Total	Learning	Physical	Visual
1. Number	2025	184	216	19	48	105
Means						
2. Actual FYA	3.48	3.40	3.38	3.49	3.46	3.31
3. Predicted FYA	3.50	3.46	3.47	3.42	3.50	3.47
4. Residual	-0.02	-0.06	-0.09	0.07	-0.04	-0.16
Mean Residuals						
5. Low Predicted	-0.06	0.10	-0.04	0.10	-0.08	0.06
6. Medium Predicted	-0.00	-0.10	-0.02	-0.31	0.12	-0.11
7. High Predicted	0.02	-0.15	-0.20	-0.22	-0.16	-0.28
Standard Deviations						
8. Actual FYA	0.42	0.50	0.52	0.48	0.55	0.54
9. Predicted FYA	0.23	0.20	0.20	0.20	0.16	0.20
10. Residuals	0.33	0.49	0.51	0.53	0.49	0.53
Correlations						
11. Actual & Predicted	0.63	0.24	0.27	0.23	-0.04	0.29

Table 3-2

Residual Analysis Derived from Predictions
Based on Test Scores Alone

	Nonhandicapped	Handicapped				
		Regular	Special			
			Total	Learning	Physical	Visual
1. Number	2025	184	216	19	48	105
Means						
2. Actual FYA	3.48	3.40	3.38	3.49	3.46	3.31
3. Predicted FYA	3.50	3.50	3.50	3.48	3.51	3.51
4. Residual	-0.02	-0.10	-0.12	0.01	-0.05	-0.20
Mean Residuals						
5. Low Predicted	-0.12	0.02	-0.07	0.02	-0.03	-0.07
6. Medium Predicted	-0.01	-0.11	-0.11	0.28	-0.03	-0.04
7. High Predicted	0.07	-0.19	-0.25	-0.39	-0.09	-0.37
Standard Deviations						
8. Actual FYA	0.42	0.50	0.52	0.48	0.56	0.54
9. Predicted FYA	0.23	0.17	0.14	0.13	0.13	0.13
10. Residuals	0.35	0.50	0.53	0.51	0.57	0.55
Correlations						
11. Actual & Predicted	0.56	0.15	0.12	-0.12	0.04	0.11

Table 3-3

Residual Analysis Derived From Predictions
Based on UGPA Only

	Nonhandicapped	Handicapped				
		Regular	Special			
			Total	Learning	Physical	Visual
1. Number	2025	184	216	19	48	105
Means						
2. Actual FYA	3.48	3.40	3.38	3.49	3.46	3.31
3. Predicted FYA	3.50	3.46	3.48	3.44	3.51	3.48
4. Residual	-0.02	-0.06	-0.10	-0.05	-0.05	-0.17
Mean Residuals						
5. Low Predicted	-0.04	0.02	-0.06	0.15	-0.12	-0.06
6. Medium Predicted	-0.02	-0.04	-0.15	0.03	-0.02	-0.30
7. High Predicted	0.01	-0.18	-0.10	-0.04	-0.06	-0.14
Standard Deviations						
8. Actual FYA	0.42	0.50	0.52	0.48	0.56	0.54
9. Predicted FYA	0.22	0.17	0.17	0.11	0.17	0.18
10. Residuals	0.34	0.49	0.50	0.48	0.52	0.53
Correlations						
11. Actual & Predicted	0.59	0.21	0.29	0.04	0.36	0.24

Table 3-4

Correlations Between Residuals and Predictors and Predicted FYA for Handicapped Students

	Predicted FYA Obtained from		
	Test Scores and UGPA	Test Scores Only	UGPA Only
Regular			
Residual-V	-.16	-.19	.04
Residual-Q	-.08	-.10	.11
Residual-U	-.10	.14	-.12
Residual-PFYA	-.15	-.18	-.13
Special-Visual			
Residual-V	-.14	-.09	-.04
Residual-Q	-.12	-.12	.00
Residual-U	-.02	.20	-.03
Residual PFYA	-.14	-.14	-.10
Special-Learning			
Residual-V	-.24	-.20	-.14
Residual-Q	-.21	-.01	-.05
Residual-U	-.13	.97	-.08
Residual-PFYA	-.36	-.37	-.19
Special-Physical			
Residual-V	.05	.09	.18
Residual-Q	-.23	-.20	-.11
Residual U	.01	.25	.04
Residual PFYA	-.08	-.19	.05

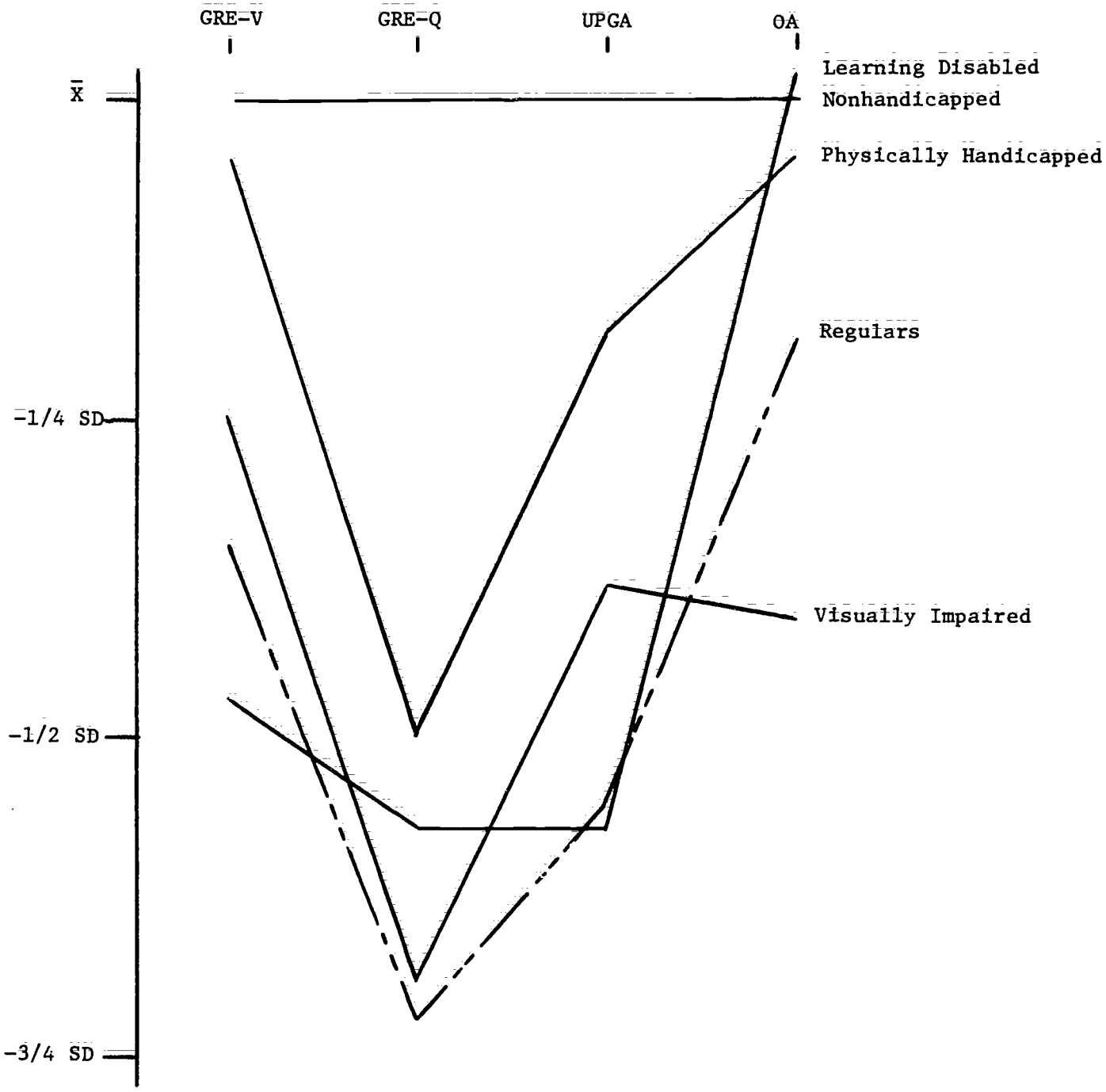

Figure 2-1. Graphical Summary of the Performance of Disabled Students Relative to the Performance of Nonhandicapped Students. (In Standard Deviation Units of the Nonhandicapped Population)

References

American Psychological Association. (1985). Standards for educational and psychological testing. Washington, D.C.: Author.

Bennett, R. E., Ragosta, M., & Stricker, L. (1984). The test performance of handicapped people (RR-84-32). Princeton, NJ: Educational Testing Service.

Braun, H. I., and Jones, D. H. (1985). Use of empirical Bayes methods in the study of the validity of academic predictors of graduate school performance. (RR-84-34). Princeton, NJ: Educational Testing Service.

Braun, H. I., Jones, D. H., Rubin, D. B., & Thayer, D. T. (1983). Empirical Bayes estimation of coefficients in the general linear model from data of deficient rank. Psychometrika, 48, 71-181.

Braun, H. I., Ragosta, M., & Kaplan, B. (1986). The predictive validity of the Scholastic Aptitude Test for disabled students. Princeton, NJ: Educational Testing Service. (in press)

Educational Testing Service. (1983) GRE 1983-84 Information Bulletin. Princeton, NJ: Author.

Livingston, S.A., & Turner, N.J. (1982). Effectiveness of the Graduate Record Examinations for predicting first-year grades: 1980-81. Summary Report of the Graduate Record Examinations Validity Study Service.

Rubin, D. B. (1980). Using empirical Bayes techniques in the law school validity studies. Journal of the American Statistical Association, 75, 801-816.

Sherman, S., & Robinson, N. (1982). Ability testing and handicapped people: Dilemma for government, science, and the public. Washington, D.C.: National Academy Press.

Wilson, K. M. (1982). A study of the validity of the restructured GRE Aptitude Test for predicting first year performance in graduate study. GREB Research Report No. 78-6R.

APPENDIX A

Empirical Bayes (EB) methods in the test validation setting have been described by Rubin (1980) and Braun and Jones (1985). In a companion paper (Braun, Ragosta and Kaplan, 1986) the authors have described how the use of EB estimation techniques facilitated the analysis of the validity of the SAT scores obtained by students with different disabilities. The purpose of this study was to carry out a similar investigation of the validity of GRE scores obtained by students with different disabilities. Unfortunately, differences in the data available in the two studies precluded simply borrowing the methodology employed in the SAT study.

Briefly, the model employed takes the form:

$$Y_{ij} = B_{0i} + B_{1i} V_{ij} + B_{2i} Q_{ij} + B_{3i} U_{ij} + e_{ij} \qquad (1)$$

where i indexes graduate departments and j indexes students within departments. The criterion, Y, is the first year average (FYA) in graduate school. V and Q represent scores on the verbal and quantitative sections of the GRE, rescaled by dividing by 200. Thus, the regression coefficients for these variables should be of comparable magnitude to that for undergraduate grade point average (UGPA), denoted by U in the equation, which is on a 0-4 scale. The errors e_{ij} are assumed to be independent and normally distributed with zero mean and variance σ_i^2.

Interest centers on estimation of the vector of parameters

$$B_i = (B_{0i}, B_{1i}, B_{2i}, B_{3i})'.$$

The empirical Bayes formulation takes the form of an hierarchical linear model by assuming in addition to (1) that

$$B_i' = Z_i'G + D_i'. \qquad (2)$$

where Z_i is a vector of department-level characteristics, G is a matrix of coefficients to be estimated and D_i is a vector of random fluctuations:

$$D_i \sim N(o, \Sigma^*). \qquad (3)$$

The model encompassed by (1), (2), and (3) facilitates the sharing of information across departments since the empirical Bayes estimate of B_i, $\tilde{B}_i$, will depend not only on the data from department i (as would the least squares estimate, $\hat{B}_i$) but also on the value of $Z_i'\tilde{G}$, a point on the plane characterized by the matrix $\tilde{G}$. ($\tilde{G}$ represents the maximum likelihood estimate of G.) All the departments contribute to the estimation of G and hence will influence the value of $\tilde{B}_i$. For more details, see Braun and Jones (1985).

On one hand, the use of EB methods seems particularly appropriate here because the unit of analysis is a graduate department with typically small enrollments, rather than a college with substantial enrollments. Consequently, the gains in efficiency of estimation over classical least squares promise to be large. On the other hand, the data available in this study is completely inadequate. Specifically, the graduate students with disabilities identified for this study were located at 261 different departments, very few of which had participated in the Validity Study Service (VSS) sponsored by the Graduate Record Examination Board. Thus there was

virtually no control data (i.e., data from nonhandicapped students) available for the estimation of the baseline equations.

One solution proposed was to use data from departments participating in the VSS to estimate the model (1) and (2), especially the matrix G. Then, if values for the components of the covariate vector Z for departments in the study could be determined, estimates of the prediction equations for those departments could be made. That is, given an estimate $\tilde{G}$ of G and a vector of covariates Z, an estimate of the coefficients of the prediction equation for the department is given by

$$\tilde{B} = Z'\tilde{G}$$

This estimate corresponds to "shrinking" the (unknown) least squares estimate for that department all the way to the plane characterized by $\tilde{G}$. However, the lack of control data implies that the values of the components of Z (usually taken to be the means of the different predictors among registered students in the department) could only be obtained with great difficulty and at substantial expense.

A more practical alternative then suggested itself; namely, to employ for the components of Z the means of the different predictors among students having their scores sent to the department. That data is relatively easily available in the GRE History files maintained at ETS. Of course this would require estimating the model for the departments in the VSS also using these new covariates and verifying that the resulting estimated prediction equations had characteristics similar to those obtained previously.

One further difficulty: with no control data from the departments in the study, it would be impossible to properly standardize the FYAs obtained by the handicapped students in order to compute usable residuals. Accordingly, it was decided to experiment with EB models using unstandardized criterion data.

For the experiments with the different versions of an EB model, we employed data on nearly 2100 students from 99 departments that participated in the VSS during the years 1980, 1981, 1982, and 1983. (The bulk of the data was collected in 1982 and 1983.) Only students that were native English speakers were retained and each department must have had at least ten students. For further details consult Swinton (1986).

From each department three students were selected at random and held over for cross-validation. The remaining students were used in the estimation process. We then fit two alternative versions of the model (1) and (2). Both versions incorporated (1) with the three predictors. However, the first version used the mean GRE-V, mean GRE-Q and mean UGPA for attending students as covariates. The second version used mean GRE-V and mean GRE-Q for score senders to the department. (UGPAs for these score-senders were unavailable.) At this stage of the investigation, standardized criterion data were still employed.

Parameter estimates were obtained using the E-M algorithm (Braun and Jones, 1985) and estimated prediction equations for each department were derived from both versions. From these equations predicted FYAs were computed for those students set aside for cross-validation. The mean squared error of prediction was used to compare the performances of the two models which, as it turned out, were very similar.

Thus encouraged, we proceeded to the second stage that was identical to the first except that the criterion, FYA, was no longer standardized. Instead, we simply scaled FYAs to fall in the range of 0-4 for all departments. Again both versions of the model were estimated and cross-validated. The results are presented below.

Three random samples of students, each containing one student from each department, were constructed from the cross-validation sample. We compared the performances of version 1 (Attending model), version 2 (Applicant model) and equations estimated by ordinary least squares using data from the particular department alone. For versions 1 and 2 the prediction equations are the usual EB estimates. Table A-1 presents five-number summaries of the absolute residuals for these approaches, separately for each of these samples. It is evident that the two versions based on EB methodology do much better than least squares estimation. Typical residuals are smaller and the maximum residuals are considerably smaller. More important for our purposes, the two versions are very similar in performance with little to choose between them.

We then compared the Attending model and the Applicant model using equations derived by shrinking down to the plane. That is, the estimates of the vector of regression coefficients under the two models for department i are given by:

$$\tilde{B}_i(1) = Z_i(1)' \tilde{G}(1), \quad \text{(Attending model)}$$

and

$$\tilde{B}_i(2) = Z_i(2)' \tilde{G}(2), \quad \text{(Applicant model)}$$

where $\tilde{G}^{(1)}$ and $\tilde{G}^{(2)}$ are estimates of the matrix of coefficients in equation (2) and $Z_i^{(1)}$ and $Z_i^{(2)}$ are the appropriate vectors of covariates. We display five-number summaries of the absolute residuals for each of the three cross-validation samples (Table A-2). Again the performances of the two models are similar, but not as good as when the two EB estimates were employed.

Nonetheless, this empirical analysis has shown that prediction equations derived from unstandardized criterion data and using applicant data to form the covariates can perform tolerably well even when control data is not employed directly. Consequently, for this study, mean test scores of applicants for each department were obtained from GRE files and combined with $\tilde{G}^{(2)}$ (estimated from the 99 departments described above) to obtain an estimate of the prediction equation for that department. Using these prediction equations, residuals for handicapped students were generated and analyzed.

Of course the derivation of these equations requires extrapolation from a model estimated from one set of departments to an entirely different set of departments. Consequently, a question concerning the accuracy of these equations arises.

What we want to do is to compare the residuals we would have obtained using empirical Bayes-estimated prediction equations employing attending student covariate data with the residuals obtained using prediction equations derived from the regression plane employing applicant data as covariates. Unfortunately, we can not do this for the 261 departments of interest. We can, however, do it for the 99 departments used to calibrate the model. We found that for each department the prediction equation generated by one model

tends to be entirely above, or entirely below, the prediction equation generated by the other. (Consequently, one set of residuals tends to be systematically more negative, or more positive, than the others.) The correlations between the two groups of predicted values are very close to unity and the typical difference in the height of the planes is quite small.

We can go a bit further. Six of the departments that contributed data to our study of handicapped students also participated in the VSS and were among the 99 departments that formed the data base for our estimation of the model parameters. There were fourteen handicapped students enrolled in these 6 departments: nine had taken the regular form of the examination, four were visually impaired and had taken a special form of the examination and one, for whom there was no information on disability, had taken a special form. Although the number of observations is rather meager, this sample provides an opportunity to compare the distribution of residuals obtained through the use of the two models described above.

Accordingly, the two sets of prediction equations for the 6 departments were generated and residuals for the students computed. The prediction equations employed students' test scores and UGPA as predictors. The two distributions have approximately the same shape, but the median residual under the reference empirical Bayes model is .12 units (approximately one-third of the interquantile range) below the median residual for the Applicant model. However, this difference is also found for nonhandicapped students in these same 6 departments.

Admittedly, the 6 departments are a small and nonrandom sample of the 261 departments and the number of students involved is less than four percent of the handicapped students in the study. Nonetheless, they provide the only link between the group of departments employed in model parameter estimation

and the group of departments employed in the residual analysis that forms the core of the study. As a further check, two analogous comparisons were run: one for the case of prediction equations incorporating only the test score as predictors and one for the case of prediction equations incorporating only UGPA as a predictor. In the two cases, the results were similar to those already described, with the median residual for the Attending model substantially more negative than the median residual for the Applicant model, both for nonhandicapped and handicapped students. Consequently, we must conclude that there is no evidence in the data that the residuals for handicapped students employed in the study are systematically biased in one direction or another.

Table A-1
Five-Number Summaries of Absolute Residuals for Three Cross-Validation Samples.

MODEL	Sample 1		Sample 2		Sample 3	
Attending	.23[c]		.18		.18	
	.09[b]	.39[d]	.08	.37	.09	.32
(EB Estimate)	0.0[a]	1.90[e]	0.0	.95	0.0	1.08
Applicant	.22		.20		.17	
	.08	.39	.07	.39	.08	.31
(EB Estimate)	0.0	1.82	0.0	.98	0.0	1.02
Least Squares	.25		.20		.22	
	.08	.48	.09	.38	.07	.41
	0.0	2.12	0.0	2.07	0.0	1.83

[c] median
[b] lower quantile
[d] upper quantile
[a] minimum
[e] maximum

Table A-2
Five-Number Summaries of Absolute Residuals for Three Cross-Validation Samples

MODEL	Sample 1		Sample 2		Sample 3	
Attending	.26		.23		.24	
	.12	.45	.11	.40	.11	.40
(EB-Plane)	0.0	1.73	0.0	1.08	0.0	1.31
Applicant	.26		.25		.24	
	.13	.44	.11	.42	.12	.41
(EB-Plane)	0.0	1.56	0.0	1.15	0.0	1.30

APPENDIX B

The following previous reports from "Studies of Admissions Testing and Handicapped People" are available upon request from Educational Testing Service, Research Publications Unit-Room T143, Princeton NJ 08541:

#1 Bennett, R., and Ragosta, M. A Research Context for Studying Admissions Tests and Handicapped Populations, 1984. (ETS Research Report 84-31)

This is the first of a series of reports emanating from a four year research effort to further knowledge of admissions testing and handicapped people. The authors describe the legal and educational issues that gave rise to this research and the major questions to be addressed. They discuss the distinguishing characteristics of different types of disability and the complex definitional problems that hamper any simple method of classifying examinees by type of handicap.

#2 Bennett, R., Ragosta, M., and Stricker, L. The Test Performance of Handicapped People, 1984 (ETS Research Report 84-32)

The purpose of this report was to summarize existing research information concerning the performance of handicapped people on admissions and other similar tests. As a group, handicapped examinees scored lower than did the nonhandicapped. Among the four major groups examined, physically handicapped and visually impaired examinees were most similar to the nondisabled population. Hearing disabled students performed least well. Available studies of the SAT and ACT generally supported the validity of those tests for handicapped people, but it was confirmed that research to date has been quite limited and has not addressed many important questions.

#3 Bennett, R., Rock, D., and Kaplan, B. The Psychometric Characteristics of the SAT for Nine Handicapped Groups, 1985. (ETS Research Report 85-49)

In this study the main finding was that with the exception of performance level, the characteristics of the Scholastic Aptitude Test (SAT) were generally comparable for handicapped and nonhandicapped students. The analyses focused on level of test performance, test reliability, speededness, and extent of unexpected differential item performance on the SAT. Visually impaired students and those with physical handicaps achieved mean scores similar to those of students taking the SAT in national administrations, while learning disabled and hearing impaired students scored lower than their nondisabled peers. Analysis of individual items revealed only a few instances of differential item performance localized to visually imparied students taking the Braille test.

#4 Rock, D., Bennett, R., and Kaplan, B. The Internal Construct Validity of the SAT Across Handicapped and Nonhandicapped Populations, 1985. (ETS Research Report 85-50)

This study further investigated the comparability of SAT Verbal and Mathematical scores for handicapped and nonhandicapped populations. A two-factor model based on Verbal and Mathematical item parcels was posed and tested for invariance across populations. This model provided a reasonable fit in all groups, with the mathematical reasoning factor generally showing a better fit than the verbal factor. Compared with the nonhandicapped population, these factors tended to be less correlated in most of the handicapped groups. This greater specificity implies the increased likelihood of achievement growth in one area independent of the other and suggests that SAT Verbal and Mathematical scores be interpreted separately rather than as an SAT composite. Finally, there was evidence that the Mathematical scores for learning disabled students taking the cassette test may underestimate the reasoning ability of this group.

#5 Ragosta, M., and Kaplan, B. A Survey of Handicapped Students Taking Special Test Administrations of the SAT and GRE, 1986 (ETS Research Report 86-5).

Disabled people were surveyed to obtain their views on the appropriateness of special test accommodations available for the Scholastic Aptitude Test (SAT) and the Graduate Record Examinations (GRE). More than nine out of ten respondents reported satisfaction with special test accommodations. A minority experienced dissatisfaction with the level of test difficulty or about specific shortcomings associated with test administrations. In comparing SAT and GRE administrations with accommodations normally provided in college testing, respondents reported that the admissions tests were more frequently offered in special versions and with extra time than were college tests.

#6 Bennett, R., Rock, D., and Jirele, T. The Psychometric Characteristics of the GRE General Test for Three Handicapped Groups, 1986. (ETS Research Report 86-6).

This study investigated four psychometric characteristics of the GRE across handicapped and nonhandicapped groups: score level, reliability, speededness, and extent of unexpected differential item performance. Results showed the performance of visually handicapped students to closely approximate that of nonhandicapped examinees, while physically handicapped students performed substantially lower. Indications of speededness were suggested for those handicapped groups taking standard as opposed to special administrations. There was no evidence of higher or lower performance on any category of items on the GRE General Test than total score would indicate, suggesting that the different item categories operate similarly for handicapped and nonhandicapped groups.

#7 Rock, D., Bennett, R., and Jirele, T. The Internal Construct Validity of the GRE General Test Across Handicapped and Nonhandicapped Populations, 1986. (ETS Research Report 86-7).

The comparability of General Test scores for handicapped and nonhandicapped groups was investigated through confirmatory factor analysis. A three factor model was posed and tested for invariance across groups. The model provided a good fit in the nonhandicapped population, a moderately good fit for visually impaired students taking the General Test under standard conditions, and the least adequate fit for visually impaired students taking the large-type edition and physically handicapped students taking the standard test. For these latter two groups, differences in internal structure were traced to the Analytical scale, whose scores appeared to have a different meaning from those for nonhandicapped students.

#8 Braun, H., Kagosta, M., and Kaplan, B. The Predictive Validity of the Scholastic Aptitude Test for Disabled Students, 1986. (ETS Research Report 86-38).

This study employed empirical Bayes procedures to determine whether prediction equations based on nonhandicapped students accurately predict first year undergraduate grades of handicapped students. Using SAT scores, previous grades, or both predictors together, college performance of handicapped students was somewhat less predictable than that of nonhandicapped students. SAT scores alone tended to overpredict college grades of learning disabled students and underpredict college grades of hearing impaired students. There was little evidence of significant over- or underprediction when grade predictions were based on both test scores and previous grades.

#9 Powers, D. E., and Willingham, W. W. Feasibility of Rescaling Test Scores of Handicapped Examinees, 1986. (ETS Research Report 86-39).

A number of testing programs offer handicapped examinees the opportunities to take admissions tests under nonstandard conditions. Typically, the test scores based on these nonstandard administrations have been flagged to indicate to test users that the scores may not be comparable to those from standard administrations. In order to avoid thus identifying handicapped applicants, a panel established by the National Academy of Sciences suggested the possibility of rescaling the scores of handicapped students to make grade predictions comparable for handicapped and nonhandicapped students. This report examines this possibility and concludes that the approach is not feasible technically and furthermore that it would also have a number of potentially serious undesirable side effects.