GMAC®

Graduate Management Admission Council®
Creating Access to Graduate Business Education®

Do Accommodated GMAT® Test Takers Have an Unfair Advantage?

Kendra Johnson, Lawrence Rudner, and Ronald Sibert

GMAC® Research Reports • RR-08-02 • July 11, 2008

Abstract

The provision of disability-related accommodations in a high-stakes testing environment raises important questions about the comparability of test scores between accommodated test takers (individuals with disabilities) and their counterparts who test under standard conditions, the validity and utility of test scores generated under accommodated conditions, and whether the accommodations themselves constitute an unfair advantage. Previous attempts to answer such questions have been primarily policy and judgment based. Sample size limitations and demographic and background differences between accommodated and non-accommodated test takers make quality empirical data difficult to gather and analyze. Drawing on a database of more than one million accommodated and non-accommodated test takers, this large-scale empirical study—the largest ever conducted to date—employed propensity score analysis to match 2,305 accommodated test takers with a comparable group of non-accommodated examinees. The key finding is that there were no meaningful or statistically significant differences in the distributions of Graduate Management Admission Test® (GMAT®) Quantitative, Verbal, Total, or writing scores for accommodated versus non-accommodated test takers when demographic and background characteristics were taken into account. Means were extremely close and effect sizes were less than .01. This all suggests that a well-designed program to provide accommodations can assure an appropriate testing environment for persons with disabilities without penalizing or giving an advantage to either group of test takers.

Introduction

Standardization, the key operating principle for achieving comparability in standardized testing, must be maintained in order for a test to be an effective evaluation tool. The integrity of the standardized construct is maintained by keeping both the test instrument and the testing conditions constant. In this way, observed score variations are most likely to reflect true individual performance differences, as opposed to measurement biases. However, for some test takers with disabilities, tests that are administered under standard timed conditions may not accurately reflect the abilities of the test taker. In such instances, the test may simply provide a measure of the nature and extent of the disability (Munger & Lloyd, 1991). On the other hand, providing an accommodation, particularly an extension of time, alters the testing conditions and could, conceivably, provide an unfair advantage.

This study examines the fairness issue by comparing GMAT scores for accommodated and non-accommodated test takers. The key methodological issue, the differing backgrounds of these two groups of test takers, is addressed through propensity score analysis.

To safeguard against measurement bias, the Americans with Disabilities Act (ADA) of 1990 requires that "when an examination is administered to an individual with a disability…the examination results accurately reflect the individual's aptitude or achievement level…rather than reflecting the individual's impaired skills" (Department of Justice [DOJ], 1996; Geisinger, 1994). In other words, federal legislation mandates

that, when necessary, reasonable test accommodations be made for test takers with disabilities. Of course, a valid and comparable measure of the student's abilities should be the outcome of the test when administered under accommodated testing conditions. Test accommodations are intended to change how a construct is being measured, not what is being measured (Stretch & Osborne, 2005). The subsequent challenge for psychometricians and educational policy makers is how to perform valid, reliable assessments of test takers with disabilities without compromising the validity of the test. In his extensive discussion on the question of validity in testing accommodations, Sireci asserts:

Four questions should be answered when determining the validity of scores from accommodated tests. These questions and the answer[s] necessary for the test to be valid include the following:

1. Does providing a particular accommodation to a particular student improve measurement of the student's knowledge, skills, and abilities? Yes.
2. Does providing a particular accommodation to some, but not all, students unfairly advantage the students who receive the accommodation? No.
3. Does providing a particular accommodation change the construct the test is measuring? No.
4. Are scores from accommodated and standard test administrations comparable? Yes. (Sireci, 2005, p.1)

Educational researchers have conducted numerous studies attempting to address one or more of the four questions raised. Sireci, Li, and Scarpati (2003) conducted the most comprehensive review of the literature to date identifying 150 test accommodation studies, of which only 28 of the 59 that focused on the effects of accommodations were found to involve empirical analysis. It was this group of 28 studies that was reviewed with respect to the interaction hypothesis. In essence, this hypothesis asserts that test accommodations, in order to be valid and fair, should improve the scores of test takers with disabilities but not improve scores of individuals for whom accommodations are not intended (Sireci, Scarpati, & Li, 2005). While the studies varied on a number of factors (e.g., accommodation types, heterogeneity of the populations studied, etc.) making generalized conclusions difficult, a consistent finding emerged. The accommodation of extended time improved the performance of *all* test takers. However, the performance of test takers with disabilities receiving accommodations was significantly greater. These findings may be viewed by some as further support for the argument that providing accommodations to some but not all students provides an unfair advantage. More recently, however, the interaction hypothesis upon which this argument (and its implications regarding the fairness of accommodations) is based has been called into question.

Two years after conducting the extensive review of the empirical literature discussed above, Sireci et al. (2005) called for a qualification of the interaction hypothesis that would involve de-emphasizing the component of the hypothesis that focuses on the impact of accommodations on non-disabled students and placing greater emphasis on the effectiveness of accommodations for students with disabilities.

> When SWD [students with disabilities] exhibit greater gains with accommodations than do their general education peers, an interaction is present. When the gains experienced by SWD are significantly greater than the gains experienced by their general education peers, the fact that the general education students achieved higher scores with an accommodation condition does not imply that the accommodation is unfair. It could imply that the test conditions are too stringent for *all* students (Sireci et al, 2005, p. 481).

In any case, understanding the effects of testing accommodations on the performance of test takers with disabilities is critical to judgments concerning the validity of the resulting scores. Of equal importance are the implications of these judgments for test takers and, as suggested earlier, for educational policy development within the testing industry. One such policy is the flagging of accommodated test scores to inform schools when a test is administered under non-standardized conditions (i.e., when an accommodation has been provided). This practice comes with an implicit presumption that the scores of accommodated students are not comparable and should be treated

differently from those of students who tested without accommodations. However, flagging scores in cases where the accommodation has no discernable impact on test validity may be tantamount to unlawful discrimination. The challenge, however, is to reliably identify such instances when a clear definition of "score comparability" is absent.

This policy of flagging non-standardized test scores was put to the test in 2000, when a GMAT test taker with a disability, Mark Breimhorst, filed suit against the Educational Testing Service (ETS), then the developer and administrator of the GMAT exam (Sireci, 2005). The implications were far reaching within the testing industry. Mr. Breimhorst contested the ETS policy of flagging scores of test takers with disabilities, and ETS ultimately settled the case by agreeing to stop flagging score reports of the GMAT exam, the Graduate Records Exam (GRE), and the Test of English as a Foreign Language (TOEFL).[1] Effective October 1, 2003, the College Board adopted a no-flagging policy for the balance of its tests administered under non-standardized (i.e., accommodated) conditions (Disability Policy Newsbreak, 2002; Sireci, 2005). American College Testing (ACT) followed suit shortly thereafter, and today several standardized testing organization policies disallow the flagging of accommodated test scores.

Guidelines for flagging test scores are enumerated in professional technical standards for the testing industry. The leading authority in this regard is the *Standards for Educational and Psychological Testing* developed jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. On the issue of flagging, the most relevant standard reads:

> *…when there is credible evidence of score comparability across regular and modified administrations, no flag should be attached to the score. When such evidence is lacking, specific information about the nature of the modification should be provided, if permitted by law, to assist test users to properly interpret and act on test scores* (Standards for Educational and Psychological Testing, Standard 10.11, in Sireci, 2005, p.5).

However, despite the impact of the Breimhorst case on current educational policy within the testing industry, and in light of the mixed conclusions reflected in the existing literature on the validity of accommodated testing, the question of whether testing accommodations constitute an unfair advantage (and public perception that it may) remains.

Methodology

Data

When candidates register to take the GMAT exam, they respond to a series of questions concerning their background and their plans. In addition to basic questions such as gender, citizenship, and race (for U.S. citizens), examinees are asked about their undergraduate major, intended graduate program (e.g. MBA, MS accounting, PhD) and when they intend to enroll. The responses to these background information questions are combined with test score and other candidate data to form the central test-taker database for the Research Department at the Graduate Management Admission Council®. Of the 1,091,869 individuals taking the GMAT exam between July 1, 2001 and March 16, 2006, 4,290 received some form of accommodation. While specifics of these accommodations were not available, detailed data were available for 2005. In that year, approximately 96% of the accommodated GMAT test takers received additional test time. The other more common accommodations included additional break time, special fonts, and special physical accommodations. Approximately 72% of the accommodated test takers received more than one accommodation.

Analysis

Because accommodation requests come primarily from United States citizens and one-third of the general population of GMAT test takers is from outside the United States, we anticipated notable differences in test-taker backgrounds. If accommodated test takers differ from non-accommodated test takers, then the simple

[1] This decision initially did not include tests owned by the College Board and administered by ETS, but the College Board eventually concurred under pressure from Disabilities Rights Advocates—a law firm representing Briemhorst and two disability-related associations Journal of Disability Policy Studies, 2002).

comparison of the groups prevalent in the 28 empirical studies reviewed by Sireci, Li, and Scarpati (2003) would not be appropriate. Sireci et al. (2005) identified two significant limitations of the group of studies they examined: the homogeneity of their samples and their small sample sizes. Our analysis addresses both issues by controlling for individual differences and utilizing large sample sizes. Finally, the subjects utilized in our study fill an additional gap identified in their review. Virtually all subjects in the 28 studies examined were K-12 test takers (Sireci et al., 2005). In contrast, the GMAT test subjects employed in this study were college-age adults.

The first step in the analysis, then, was to document the extent to which accommodated test takers differ from the general population in terms of gender, race, citizenship, and other variables. The rationale was that once differences could be identified, they could be statistically isolated and their possible confounding effects eliminated. A matched-pairs sample of accommodated and non-accommodated test takers was formed using propensity score analysis (Rosenbaum & Rubin, 1985; Rubin, 1997; Joffe & Rosenbaum, 1999). With propensity score analysis, covariates are combined to yield the likelihood of a subject belonging to a given group. Accommodated individuals were then matched, based on their propensity score, to individuals in the non-accommodated group. By this process, we weighted the variables by their relative importance and matched based on an optimal composite. Rubin (1997) has shown that when one matches on the composite propensity score, the group means and standard deviations of the covariates will also be equivalent.

The dependent variables for this analysis were GMAT Verbal Scaled Score, GMAT Quantitative Scaled Score, and GMAT Total Scaled Score. It was felt that mean score differences of 1.5 points for Quant or Verbal and 15 points for Total would have practical significance.

The covariates for the analysis are listed in Table 1.

Table 1. Covariates for Analysis

Covariate Name	Description
UGPA	self-reported undergraduate grade point average
Plan Go FT	a binary variable indicating whether the examinee intends to attend a full-time program (coded 1) or not (coded 0)
Pursue MBA	a binary variable to indicate whether the examinee intended to enroll in an MBA program (coded 1) or not (coded 0). Non-MBA programs include doctoral programs and various non-MBA master level programs, e.g. Masters of Accountancy.
Age	age on day of testing
Male	a binary variable coded 1 for male andcoded 0 for female
Biz Undergrad	a binary variable to indicate whether the examinee held a business undergraduate degree
White	a binary variable coded 1 for White and 0 for all others. Race is a self report variable only asked of US citizens.
Days to Enroll-	the difference between the intended program start date and the test date
USA Citizen-	a binary variable coded 1 for citizen of the United States and 0 for all others

Results

The first analysis was based on a random sample of 15,000 non-accommodated and 2,305 accommodated test takers with complete data. As shown in Table 2, we were correct to anticipate notable differences in the two groups. The percentages of test takers who plan to enroll as full-time students, are white, are male, and are United States citizens are much higher for accommodated test takers. Accommodated test takers also tend to be slightly younger and tend to take the GMAT exam earlier. With the exceptions of the percentage of business undergraduates and undergraduate grade point average, there are

significant differences between the non-accommodated and the accommodated test takers on all of the means and proportions in Table 2 when evaluated using t-tests at $p < .05$.

Table 2: Characteristics of Unmatched Accommodated and Non-Accommodated GMAT® Test Takers

	Non-Accommodated		Accommodated		
Covariate Name	**Percentage**	**sd**	**Percentage**	**sd**	**Effect size**
Plan Go FT	60.6%	.489	75.3%	.431	0.31
Pursue MBA	79.3%	.405	82.7%	.378	0.08
White	39.9%	.490	64.4%	.490	0.50
Male	61.0%	.488	72.2%	.448	0.23
Biz Undergrad	44.4%	.497	42.0%	.493	-0.05
US Citizen	58.1%	.493	86.7%	.340	0.60
Covariate Name	**Mean**	**sd**	**Mean**	**sd**	**Effect size**
Age (yrs.)	28.19	6.32	27.29	5.07	-0.15
UGPA	3.20	.50	3.19	.46	-0.03
Days to Enroll	209.20	213.60	241.60	256.20	0.15

Only 2,305 of the 4,290 accommodated test takers had complete data on all of the covariates. In order to determine whether list-wise deletion would bias the sample, the percentages and means for the 2,305 test takers were compared against the means for all 4,290 accommodated students. T-tests found no significant differences at $p < .05$. All the means and percentages were extremely close.

Discriminant Function Analysis was used to compute propensity scores—the likelihood of having received an accommodation—as a function of the nine variables using the sample of 15,000 non-accommodated and 2,305 accommodated test takers. The canonical correlation was significantly different than zero ($r = .28$; Wilk's lambda = .922; $df = 7$; $p < .05$). The propensity score was then computed for all test takers. Each of the 2,305 accommodated test takers was matched with a randomly drawn non-accommodated test taker with the same propensity score.

Table 3 reveals that the resultant groups were matched quite well. There are no meaningful or statistically significant differences between the matched groups of accommodated and non-accommodated test takers on any of the nine variables.

Table 3: Characteristics of Matched Accommodated and Non-Accommodated GMAT® Test Takers

	Non-Accommodated		Accommodated		
Covariate Name	**Percentage**	**sd**	**Percentage**	**sd**	**Effect size**
Plan Go FT	77.3%	.419	75.3%	.431	-0.05
Pursue MBA	80.6%	.395	82.7%	.378	0.05
White	63.4%	.482	64.4%	.479	0.02
Male	72.4%	.447	72.1%	.449	-0.01
Biz Undergrad	41.8%	.493	41.9%	.493	0.00
US Citizen	85.2%	.355	86.7%	.340	0.04

Table 3: Characteristics of Matched Accommodated and Non-Accommodated GMAT® Test Takers					
	Non-accommodated		Accommodated		
Covariate Name	Mean	sd	Mean	sd	Effect size
Age (yrs.)	27.06	5.52	27.29	5.07	0.04
UGPA	3.20	.46	3.19	.45	-0.03
Days to Enroll	245.80	233.10	241.60	256.20	-0.02

The key question being explored here is whether accommodated test takers score higher than non-accommodated test takers after controlling for background differences. As shown in Table 4, the mean scores for the 2,305 accommodated test takers and those for the matched group of 2,305 non-accommodated examinees are virtually identical. None of the differences in the means are statistically or practically significant.

Table 4: GMAT® scores for matched groups of non-accommodated and accommodated test takers.					
	Non-accommodated		Accommodated		
Scores	Mean	sd	Mean	sd	Effect size
GMAT® Verbal	30.3	8.4	30.4	8.2	0.01
GMAT® Quant	34.5	9.6	34.6	9.5	0.01
GMAT® Total	544.8	112.5	546.1	113.1	0.01

Discussion

Accommodated test takers as a group differ from non-accommodated test takers along a number of important demographic dimensions and other dimensions not related to test structure or performance—most notably the percentages of test takers who plan to enroll as full-time students, are white, are male, and are U.S. citizens. When these and other background differences are taken into account (i.e., controlled), the GMAT scores of accommodated and non-accommodated test takers are virtually identical. In other words, when we selected a group of non-accommodated test takers who were similar to the accommodated test takers on select variables, their scores were almost exactly the same as the scores of non-accommodated test takers. By contrast, had we not controlled for the select variables and simply compared accommodated to non-accommodated test takers, we would have drawn a radically different, and erroneous, conclusion. The mean scores for the non-accommodated test takers are 27.2, 35.2, and 526.2 for the GMAT Verbal, Quantitative, and Total scores, respectively. This would suggest a meaningful 3-point advantage in Verbal and a 20-point advantage in Total scores. But, again, that would not have been an appropriate comparison.

In essence, the approach taken here represents a departure in focus from that of comparability analyses conducted to date, which have focused overwhelmingly on predictive validity, construct equivalence, and/or other test-related considerations (see Sireci, 2005). In contrast, this study examined characteristics of the test takers themselves, and these characteristics were treated as covariates in a performance analysis. That is, controlling for the selected covariates ostensibly removed their confounding influence, thereby allowing the analysis to focus on performance variances between comparable groups of standard and accommodated test takers. In addition, when this approach was applied, the GMAT

exam fared favorably with respect to the four validity-related questions advanced by Sireci (2005) as a means for determining the validity of scores from accommodated tests.

Implications for Education Policy Makers

There is still much debate about the fairness of providing accommodations for qualified standardized test takers with disabilities. The issue of flagging scores of disabled test takers, which may still occur under certain conditions, remains a flashpoint as well because prevailing industry standards and policy guidance remains somewhat vague about how to discern these conditions. Accommodations, therefore, still pose a tremendous challenge to policymakers and test providers. The complexity of the topic is captured in the following quote:

> The correct decision on whether to flag or not depends on the nature of the accommodation, the degree to which the accommodation alters the construct measured, and the degree to which the accommodation affects the interpretation given to a test score (Geisinger, 1994; Green & Sireci, 1999; Pitoniak & Royer, 2001). Thus, just as validity must be interpreted with respect to a given testing purpose, the appropriateness of flagging scores from accommodated tests must be interpreted with respect to the degree to which scores from the accommodated test administration are comparable to scores from standard administrations (Sireci, 2005, p.9).

The sheer variety of disabilities and their degrees of manifestation suggest not only the need for variety in the types of accommodations offered but possible variations in the effectiveness of those accommodations that are selected. From this perspective, reliable case-by-case determinations of suitability, validity, and comparability of accommodated test conditions to standard testing conditions would be prohibitively complex and cumbersome. The same might be said of the challenge associated with designing a research model that could effectively guide formulation of reliable accommodation-related standards.

However, as this study demonstrates, an examination of comparability from the perspective of *performance outcomes* (vs. characteristics of the test or the construct relevance of accommodations) may offer some measure of utility in that regard. For typical forms of accommodation, such as extended time, where there can be reasonable certainty of construct neutrality, well-executed outcome- or performance-related comparative research can effectively inform policy without the need to address complex variations in disability or corresponding types of accommodation. Our findings with respect to the comparability of scores of accommodated and non-accommodated GMAT test takers suggest that such an analysis would be particularly effective when subject sample sizes are sufficiently large to be representative of the variety of accommodations provided to the population of test takers with disabilities. The corresponding standard then might involve subjecting high-stakes tests to this form of scrutiny as a broad comparability screen for accommodated test conditions.

While this study employed the largest single dataset to date, care should be exerted not to over-generalize the findings. Its demonstrated lack of difference between the accommodated and non-accommodated groups may be attributed to the nature of the accommodation, the particular assessment, and the way the accommodated group was formed. Though the results will likely generalize to comparable high-quality assessment and accommodation evaluation procedures, the question of broader applicability will and should remain an empirical one.

Finally, the accommodations review process must evaluate each individual case based on the clinically discernable features of the particular disability that prompted the request for accommodations and the legitimacy of that request. The group findings presented here are not meant to extend to individual cases and do not relieve testing companies of the burden of a thorough review of each request.

Contact Information

For questions or comments regarding study findings, methodology or data, please contact the GMAC® Research and Development department at research@gmac.com.

References

Americans with Disabilities Act of 1990, 42 U.S.C. 91 12101 et seq (1990).

Geisinger, K. F. (1994). Psychometric issues in testing students with disabilities. *Applied Measurement in Education, 7*, 121 -140.

Joffe, M.M. and P.R. Rosenbaum (1999). Propensity Scores, *American Journal of Epidemiology*, 150, 327-333

Journal of Disability Policy Studies. (2002). Disability Policy Newsbreak!. *Journal of Disability Policy Studies*, 13(3), 190.

Munger, G. F., & Lloyd, B.H. (1991). Effect of speededness on test performance of handicapped and nonhandicapped examinees. *Journal of Educational Research,*85, 53-58.

Rosenbaum P.R. and D.B. Rubin (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician.* 39(1), 33-38

Rubin, D.B. (1997). Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Annals of Internal Medicine.* 127(8S), Supplement, 757-763

Sireci, S.G. (2004, February). Validity issues in accommodating NAEP reading tests. *Center for Educational Assessment research report no. 515*. Amherst, MA: Center for Educational Assessment, University of Massachusetts. Commissioned by Educational Testing Service.Sireci, S.G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Researcher*, 34(1), 3-12.

Sireci, S.G., Li, S., & Scarpati, S. (2003). The effects of test accommodation on test performance: A review of the literature. Commissioned paper by the National Academy of Sciences/National Research Council's Board on Testing and Assessment, Washington, DC: National Research Council.

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457-490.

Stretch, L. S., & Osborne, J. (2005). Extended time test accommodations: Directions for future research and practice. *Practical Assessment & Evaluation,*10 (8).