GMAC®

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

Exploring Potential Quantitative Item Bias across Groups

Kara M. Owens & Lawrence M. Rudner *GMAC® Research Reports • RR-07-03 • January 19, 2007*

Abstract

This study used a differential item impact (DII) procedure to flag items on the Graduate Management Admission Test® exam that might lead to bias against subgroups of examinees. Items were flagged if they resulted in different conditional probabilities of a correct response for a subgroup of U.S. or non-U.S. citizens when compared to the calibration sample of examinees with similar ability. A small set of non-operational items were identified as having relatively high DII values. Specifically, quantitative items with a reading comprehension component yielded negative impact values for non-U.S. subgroups and positive values for U.S. subgroups. The opposite pattern was revealed when purely computational quantitative items were examined. Examples of these items are provided and reasons for potential bias are explored.

With the public's and industry's shared concern for fairness in testing, it has become increasingly important to provide evidence that instruments allow for valid inferences regarding specified purposes for any potential subgroup for whom the test is administered. With an entire chapter devoted to fairness, the *Standards for Educational and Psychological Testing* addresses many of the issues involved in testing individuals from different backgrounds (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). In today's multicultural society and cross-cultural use of standardized instruments, there is a greater need for evidence investigating the validity of test inferences across various nationalities and cultures. Inferences based on test scores must be equally valid for diverse situations, people, and places. This can be especially important for admissions tests. Because many undergraduate and graduate institutions require students from around the world to take admissions tests, it is critical that the development of these tests fully considers different test-taking population subgroups.

The Graduate Management Admission Test® (GMAT®) exam, originally known as the Admission Test for Graduate Study in Business, has been in existence since 1954 (GMAC®, 1999). The GMAT® exam measures verbal, quantitative, and analytical writing skills using three separate sections. Examinees receive scores for each section as well as a Total score that combines information from both the Verbal and Quantitative sections. Continual research is conducted to validate the inferences made from GMAT® test scores to determine that the various sections accurately predict first year performance in graduate management education programs (GMAC®, 1999). A recent meta-analysis of data from more than 40,000 students showed that validity estimates for the GMAT® exam indicated the test allows for valid inferences to be made across a range of programs (Talento-Miller & Rudner, 2005). The correlation between first-year grades and GMAT® Total and analytic writing assessment (AWA) scores was .52. When undergraduate grade point average was combined with GMAT® Total and AWA scores, the multiple correlation increased to .53.

Research on the GMAT® exam has investigated predictive validity for different subgroups (Crooks & Heuvelmans, 1999; Dobson, Krapljan-Barr, & Vielba, 1999; Koys, 2005; Talento-Miller, 2005; Wilson, 1985). Several of these studies indicated that the GMAT® test provides valid scores for subgroups other than U.S. citizens (Crooks & Heuvelmans, 1999; Gallagher, Bridgeman, & Cahalan, 2002; Koys, 2005; Talento-Miller, 2005; Wilson, 1985). However, none of these studies examined the role of potential item bias in determining the validity of test score inferences for certain subgroups of the population. Specifically, studies controlling for the influence of prior ability on GMAT® exam performance and comparing performance for subgroups of non-U.S. test takers have not been made publicly available.

The present study investigated potential bias by examining performance differences among test items for United States (U.S.) and non-U.S. citizens, while simultaneously controlling for prior ability. Whereas previous research typically used U.S. citizens or White examinees as the reference group for comparisons, the current study used the calibration sample as the reference group. This methodology, referred to as differential item impact (DII; Guo, Rudner, Talento-Miller, 2006), is discussed. Items displaying potential DII using this technique are identified, and explanations are hypothesized.

Methods

Data

This analysis was based on responses to non-operational items embedded within administrations of the GMAT® exam during the 2003 and 2004 calendar years. While the operational items of the GMAT® exam are administered using computer adaptive testing technology, the non-operational items are randomly assigned to examinees. These non-operational items are organized in pretest pools of 7 to 14 items. Each examinee is randomly assigned one non-operational pool of Quantitative items and one non-operational pool of Verbal items. The items within each non-operational pool are then administered in a random order interspersed among the operational items. Each non-operational item is seen by approximately 1,400 test takers, including 800 U.S. citizens and 600 non-U.S. citizens. For this study, a sample of 1,625 non-operational items were examined for potential bias.

Analysis

Each of the 1,625 non-operational items was calibrated based on the theta estimates for all examinees and their dichotomously scored responses to the given item. Should the non-operational item become operational in the future, these item parameters would become the operational item parameters. In this paper, the parameters from this calibration are based on all test-takers and are referred to as the operational item parameters, denoted as $a_{g|o}$, $b_{g|o}$ and $c_{g|o}$, where g indicates the item number and o indicates that it is operational.

Examinees were also separated into two independent groups based on citizenship. Each non-operational item was then calibrated again, but this time the calibration was performed separately for U.S. citizens and non-U.S. citizens. These parameters are referred to as the focal group item parameters and are denoted as $a_{g|f}$, $b_{g|f}$ and $c_{g|f}$, where f indicates that it is based on the focal group. Item bias was defined as the impact of using the operational parameters rather than the focal group parameters:

$$I_{g|f} = \frac{\sum_{i \in f}^{N} \left[P\left(u_g = 1 \mid \theta_i, a_{g|f}, b_{g|f}, c_{g|f}\right) - P\left(u_g = 1 \mid \theta_i, a_{g|o}, b_{g|o}, c_{g|o}\right) \right]}{N}$$

The impact for item g and focal group f is defined at GMAC® as the mean difference for focal group examinees in the probability of a correct response based on the focal group item parameters and the probability based on the operational item parameters. This statistic offers several practical advantages:

1. It uses operational parameters, rather than another subgroup, as a baseline. The method allows for interpretations regarding how the use of actual operational parameters, rather than those estimated specifically for the subgroup of interest, will influence their scores.
2. It is self-weighted by the distribution of examinees along the theta or ability scale for the focal and operational groups.
3. The directionality of impact is easy to detect.
4. The statistic is not affected by sample size for the focal or operational groups.

A 7% difference in either direction for the focal group and operational item parameters was used as a flag to define impact. Thus, for a flagged item, the average probability of a correct response by the focal and operational groups for that item must differ by at least an absolute value of .07. Negative impact values indicate that the focal group had a lower probability than the operational group of responding correctly to the item. However, because the operational parameters were used to estimate focal group performance on the item, focal group performance was actually over-estimated on any items with negative impact values. On the other hand, positive impact values reveal that the focal group had a higher probability of a correct response than the operational group for that specific item. And because operational parameters were used to estimate the focal group's ability, rather than the focal group parameters, focal group ability was under-estimated for any items with positive impact values. Item parameters used to calculate impact for the U.S. and non-U.S. focal groups included responses from 294 to 942 examinees for each item. For the operational group, item parameters were based on between 987 and 1,796 examinee responses for each item.

The purpose of the study was to identify characteristics that might be common across flagged items. Each Quantitative item was classified as containing three of the following seven characteristics among three dimensions:

1. Higher-order skill type
    a. Data sufficiency requiring the test taker to identify whether there is adequate information to answer a problem, or
    b. Problem solving
2. Basic skill format
    a. Algebra,
    b. Geometry, or
    c. Arithmetic functions
3. Structure
    a. Applied, involving scenarios, or
    b. Pure, involving mostly equations and formulas

Each verbal item was simply classified as reading comprehension, sentence correction, or critical reasoning. Flagged Verbal and Quantitative items were then collectively examined for patterns in item characteristics.

Results

The seven characteristics (data sufficiency vs. problem solving; algebra vs. geometry vs. arithmetic functions; applied vs. pure), two groups (U.S. and non-U.S.), and 1,625 items resulted in approximately 17,000 comparisons. A total of 39 of the 1,625 non-operational Quantitative and Verbal items were identified as demonstrating at least a 7% difference in focal and operational probabilities of a correct response. Of these 39 items, 25 were from the Quantitative section of the GMAT® exam. Additionally, seven of the 25 potentially biased items from the Quantitative section demonstrated impact for both citizenship subgroups. Because of the lack of DII for the Verbal items, this study focused solely on the impact results for the handful of non-operational GMAT® Quantitative items displaying DII.

GMAT® Quantitative Section Impact

Quantitative Item Formats

The GMAT® Quantitative section requires knowledge in areas such as arithmetic, algebra, and geometry. Items in the GMAT® Quantitative section are multiple choice and measure higher-order skills including problem solving and data sufficiency. Problem-solving items are standard multiple choice questions testing mathematical skills and competency with quantitative concepts. Data sufficiency questions require examinees to determine if enough information is presented to allow them to draw specified conclusions or answer provided questions. Figure 1 presents an example of a data sufficiency question obtained from the Web site examinees use to register for the GMAT® exam.

Figure 1. Example of a Data Sufficiency Question, from www.mba.com

If a real estate agent received a commission of 6 percent of the selling price of a certain house, what was the selling price of the house?

(1) The selling price minus the real estate agent's commission was $84,600.
(2) The selling price was 250 percent of the original purchase price of $36,000.

(A) Statement (1) ALONE is sufficient, but statement (2) alone is not sufficient.

(B) Statement (2) ALONE is sufficient, but statement (1) alone is not sufficient.

(C) BOTH statements TOGETHER are sufficient, but NEITHER statement ALONE is sufficient.

(D) EACH statement ALONE is sufficient.

(E) Statements (1) and (2) TOGETHER are NOT sufficient.

Items in the GMAT® Quantitative section are also classified based on the format in which skills are being tested. For example, "real" items require a practical application of computational skills. Many items classified into this category present a situation that requires the examinee to read a word problem or evaluate a chart or graph and use the information provided to derive the correct answer. Conversely, a "pure" item solely requires computational skills to respond correctly to the item. There are often equations and formulas with very limited text in this format. Figures 2 and 3 provide examples of publicly available real and pure items, respectively.

Figure 2. Example of a "Real" Question, from *The Official Guide for GMAT® Review* (2003)[1]

In a weight-lifting competition, the total weight of Joe's two lifts was 750 pounds. If twice the weight of his first lift was 300 pounds more than the weight of his second lift, what was the weight, in pounds, of his first lift?

(A) 225

(B) 275

(C) 325

(D) 350

(E) 400

[1] Figure 2 is a Quantitative item published in *The Official Guide for GMAT® Review* (p. 76), by Educational Testing Service® and Graduate Management Admission Council®, 2003, McLean, VA: Graduate Management Admission Council®. Copyright 2003 by the Graduate Management Admission Council®. Reprinted with permission.

Figure 3. Example of a "Pure" Question, from www.mba.com

If $u > t$, $r > q$, $s > t$, and $t > r$, which of the following must be true?

(I) $u > s$
(II) $s > q$
(III) $u > r$

(A) I only
(B) II only
(C) III only
(D) I and II
(E) II and III

Summary of Quantitative Impact

Table 1 provides a summary table of positive and negative impact by citizenship group and item format. Of the set of 25 potentially biased Quantitative items, impact values above the cutoff were found for 24 items for the non-U.S. focal group. Of these 24 items, 17 demonstrated positive impact values for non-U.S. citizens, and seven items yielded negative impact values for this subgroup. Conversely, only eight of the 25 potentially biased Quantitative items were flagged for the U.S. citizen subgroup. Of these eight items, three demonstrated a positive impact value for U.S. citizens, and five revealed a negative direction of impact.

Table 1. Impact by Item Type and Subgroup

	Non-U.S.	U.S.
Positive Impact		
Pure	16	0
Real	1	3
Total Positive Items	17	3
Negative Impact		
Pure	0	5
Real	7	0
Total Negative Items	7	5
Total Items with Impact	24	8

**Note.* A total of 25 Quantitative items demonstrated potential bias. Seven of these items were biased for both non-U.S. and U.S. citizens.

Nine of the 25 items demonstrating potential bias for U.S. and/or non-U.S. citizens were classified as real, and 16 items were classified as pure. Table 2 provides a breakdown of pure and real items and their impact values based on subgroup membership. Absolute impact values deemed practically significant for this study range from a low of 0.0760 to a high of 0.1288. Thus, there was an 8-13% difference in probabilities for correct responses between the subgroups and the calibration sample for these 25 flagged items.

Table 2. Potentially Biased Quantitative Items						
			Non-U.S.		U.S.	
Item #	Skill	Structure	Positive Impact	Negative Impact	Positive Impact	Negative Impact
1	Algebra	Pure	0.0811	—	—	—
2	Algebra	Pure	0.1288	—	—	–0.0945
3	Arithmetic	Real	—	–0.0836	—	—
4	Algebra	Pure	0.0969	—	—	–0.0774
5	Arithmetic	Pure	0.0811	—	—	—
6	Algebra	Pure	0.0856	—	—	—
7	Arithmetic	Real	—	–0.0936	0.0804	—
8	Geometry	Pure	0.0935	—	—	—
9	Algebra	Pure	0.1106	—	—	—
10	Arithmetic	Pure	0.0785	—	—	—
11	Arithmetic	Real	—	–0.0789	—	—
12	Arithmetic	Real	—	–0.0795	—	—
13	Algebra	Pure	0.0841	—	—	—
14	Arithmetic	Real	—	–0.0795	—	—
15	Arithmetic	Pure	0.1109	—	—	—
16	Arithmetic	Real	—	–0.0783	0.0807	—
17	Arithmetic	Pure	0.0764	—	—	—
18	Arithmetic	Pure	0.0760	—	—	—
19	Arithmetic	Pure	0.0905	—	—	–0.0777
20	Algebra	Pure	0.0949	—	—	–0.0809
21	Algebra	Real	0.0767	—	—	—
22	Arithmetic	Real	—	–0.0826	—	—
23	Algebra	Pure	0.0955	—	—	–0.0813
24	Algebra	Pure	0.0770	—	—	—
25	Arithmetic	Real	—	—	0.0833	—

Of the 16 pure GMAT® Quantitative items flagged as potentially biased, all demonstrated positive impact in favor of non-U.S. citizens. This indicates that the non-U.S. subgroup had a higher probability than the operational sample of responding correctly to those items. Five of these pure items also yielded negative impact values for U.S. citizens. Conversely, seven of the nine potentially biased real GMAT® Quantitative items yielded negative impact values for non-U.S. citizens. All seven of these items were also classified as requiring arithmetic reasoning skills, as opposed to algebra or geometry. Counter to the negative impact values found for pure items for the U.S. subgroup, three of the potentially biased real items demonstrated a positive impact in favor of U.S. citizens.

Ultimately, non-U.S. citizens appeared to have a higher probability of responding correctly to purely computational items when compared to the operational sample. On the other hand, U.S. citizens had a higher probability than the operational sample of correctly answering word problems requiring realistic applications of quantitative skills. Additionally, when non-U.S. citizens responded to real items and U.S. citizens responded to pure items, an opposite pattern emerged. Non-U.S. citizens were less likely to respond correctly when faced with real items, and U.S. citizens had a lower probability of responding correctly to pure items. Thus, there were distinctive differences in the probability of correct responses to specific items depending on citizenship subgroup and the format of the Quantitative item administered.

It is important to note the effect of operationally calibrated parameters on focal group examinee scores for those items flagged using the impact statistic. For instance, negative impact values revealed for non-U.S. citizens on real items and for U.S. citizens on pure items indicated that the focal group parameter estimates were lower than the operational group parameters for the items. Thus, if the operational rather than focal group parameters are used to estimate theta for these items, focal group examinees would have higher estimates of ability on the items than they actually should. However, ability estimates would be under-predicted for those items in which the focal groups had a higher probability of correct responses when compared to the operational sample (i.e., positive DII value). An examination of Tables 1 and 2 in light of this note reveals that if operational parameters are used to estimate ability for the flagged items, the ability of non-U.S. citizens would be under-predicted more than over-predicted, and the opposite would be true for U.S. citizens.

U.S. Positive Impact and Non-U.S. Negative Impact Items

Although the items studied for this research were non-operational, they may still be modified and used in future item pools. Thus, actual items examined in the study will not be divulged in this paper. Instead, Figures 4 and 5 provide similar examples taken from *The Official Guide for GMAT® Review* (Educational Testing Service® & GMAC®, 2003) of some of the non-operational items flagged in this analysis. Figure 4 would be classified as a real item because it uses a practical application of quantitative skills for a specified scenario. A similar non-operational item was found to have a negative impact for non-U.S. examinees and a positive impact for U.S. examinees.

Figure 4. Example of a "Real" Quantitative Item from *The Official Guide for GMAT® Review*[2]

The arithmetic mean and standard deviation of a certain normal distribution are 13.5 and 1.5, respectively. What value is exactly 2 standard deviations less than the mean?

(A) 10.5

(B) 11.0

(C) 11.5

(D) 12.0

(E) 12.5

[2] Figure 4 is a *Quantitative item* published in *The Official Guide for GMAT® Review* (p. 123), by Educational Testing Service® and Graduate Management Admission Council®, 2003, McLean, VA: Graduate Management Admission Council®. Copyright 2003 by the Graduate Management Admission Council®. Reprinted with permission.

This item requires examinees to demonstrate arithmetic ability with standard deviation. When reviewing other items that revealed negative impact values for non-U.S. citizens, all seven negative impact items required arithmetic skills rather than algebra or geometry ability. Items requiring skills in calculating ranges and selecting values within or outside of these ranges appeared to be especially problematic for non-U.S. examinees when compared to the operational sample.

Perhaps non-U.S. students experience difficulty interpreting the GMAT® Quantitative items when they are written in the "real" format, as opposed to the "pure" format. Non-U.S. examinees are also likely reading these items in a language that is secondary to them. Furthermore, countries outside the U.S. may not focus as much on basic arithmetic skills. Based on the items revealing positive impact values for U.S. and negative impact values for non-U.S. examinees, there appears to be a discrepancy in arithmetic and reading comprehension skills for the two subgroups when compared to the calibration sample.

Non-U.S. Positive Impact and U.S. Negative Impact Items

Figure 5 represents an example of a pure item on which U.S. citizens had a low probability of responding correctly and non-U.S. citizens had a high probability of responding correctly. Pure items such as this one require examinees to use formulas and algebraic skills with positive and negative exponents to derive the correct answer. In fact, four of the five items demonstrating negative impact for U.S. citizens required algebra skills. Of those, three required competency with negative exponents. While U.S. examinees were more likely than the calibration sample to respond incorrectly to the algebra items, the opposite was true for non-U.S. examinees on these items. Thus, it appears that non-U.S. citizens have a much greater capacity for algebraic computations, especially those involving negative exponents, than their U.S. counterparts. U.S. examinees, however, outperformed the calibration sample on three arithmetic items, based on the positive impact values calculated for U.S. examinees on these items. Perhaps U.S. examinees are more skilled in the area of arithmetic than they are in algebraic computations.

Figure 5. Example of a "Pure" Quantitative Item from *The Official Guide for GMAT® Review*[3]

If m is an integer such that $(-2)^{2m} = 2^{9-m}$, then $m =$

(A) 1

(B) 2

(C) 3

(D) 4

(E) 6

[3] Figure 5 is a *Quantitative item* published in *The Official Guide for GMAT® Review* (p. 82), by Educational Testing Service® and Graduate Management Admission Council®, 2003, McLean, VA: Graduate Management Admission Council®. Copyright 2003 by the Graduate Management Admission Council®. Reprinted with permission.

Based on these examples and a review of Table 2, the findings may offer evidence that the U.S. educational system focuses more attention on practical arithmetic skills and less on algebra proficiency and computational ability. Conversely, educational systems outside of the U.S. may emphasize competence with algebraic and arithmetic formulas and calculations, as evidenced by the positive impact values reported for items in both of these domains. At the same time, proficiency may be lacking for non-U.S. citizens in the area of practically applied arithmetic functions, such as skills with word problems.

Conclusion

This study examined items from the GMAT® exam for potential bias among U.S. and non-U.S. examinees. By selecting a method to identify DII that allowed for focal group comparisons with operational item parameters, different and more meaningful information can be gathered than that which has been revealed by previous research. Additionally, the impact value estimate used in this study was not influenced by sample size and allowed for an easier investigation of directionality of potential DII for the subgroups. Calculations of impact yielded 25 Quantitative and 14 Verbal items with impact values above a cutoff value of .07 from a pool of 1,625 items. Thus, for these items, there was at least a 7% difference in the probability of a correct response for the subgroup when compared with the operational item parameters. Though being flagged does not necessarily mean that items are functioning differently for the diverse subgroups, it does warrant further investigation of the flagged items. Items could also be flagged because of sampling or model fit errors, rather than as a result of true differences between subgroup and operational item parameters. However, when the flagged items were investigated further, a clear pattern, based on item format and content, emerged as a distinguishing factor influencing performance on these items among U.S. and non-U.S. citizens.

By comparing Figures 4 and 5, it becomes obvious that there were distinct differences in subgroup performance between items presented in "real" versus "pure" formats and items requiring arithmetic versus algebra skills. Though non-U.S. citizens had a higher probability of responding correctly to a set of "pure" GMAT® Quantitative items, they were also less likely to respond correctly to items that included practical applications of arithmetic skills. On the other hand, U.S. citizens had a higher probability of identifying correct responses on these "real" GMAT® Quantitative items but a lower probability of selecting a correct response for the purely computational arithmetic and algebra items.

It is important to note the implications of using the operational item parameters to estimate ability for subgroups, which is standard practice for most testing programs. When a subgroup demonstrates a higher probability of responding correctly to an item than the operational group (i.e., positive DII values), the subgroup's ability estimate for that specific item will be underestimated when the operational item parameters are used. Conversely, if the subgroup has a lower probability of correctly responding to a specific item than the operational sample (i.e., negative DII values), the subgroup's ability for that item will be over-estimated if operational parameters are used.

Caution is warranted when trying to distinguish the potentially biasing factor in the items used in this analysis. It was difficult to determine if the biasing factor was the item format, a unique aspect of the item content, or some hidden feature in the item text that was influencing discrepancy in U.S. and non-U.S. performance. Very few items were flagged of the 17,000 comparisons. Thus, generalizing from scant examples is tenuous at best.

Though the results are far from definitive, they are suggestive. Further investigation, specifically targeting the flagged item types, might result in improved guidelines for cross-cultural item development. One such study might involve modifying and re-testing these items to determine features that might be altered to reduce the impact values. Additionally, future research should seek to compare U.S. and non-U.S. educational systems and preparation practices for quantitative skills and abilities. Though this study examined GMAT® items and GMAT® test takers exclusively, an investigation cross validating the results with other examinations could prove to be quite informative.

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (1999). *Standards for Educational and Psychological Testing.* Washington, D. C.: American Educational Research Association.

Crooks, S., & Heuvelmans, A. (1999, October). *The GMAT® as a predictor of academic performance on eight European MBA programs.* Arnhem, The Netherlands: The National Institute for Educational Measurement.

Dobson, P., Krapljan-Barr, P., & Vielba, C. (1999). An evaluation of the validity and fairness of the graduate management admissions test (GMAT®) used for MBA selection in a UK business school. *International Journal of Selection and Assessment, 7*(4), 196-202.

Educational Testing Service®, & Graduate Management Admission Council® (Eds.). (2003). *The official guide for GMAT® review.* McLean, VA: GMAC®.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. *Journal of Educational Measurement, 39*(2), 133-147.

Graduate Management Admission Council®. (1999). *Guide to the use of GMAT® scores.* [Brochure]. Princeton, NJ: Author.

Guo, F., Rudner, L. M., & Talento-Miller, E. (2006, July). *Differential impact as an item bias indicator in CAT and other IRT-based tests.* Paper presented at the annual meeting of the International Test Commission, Brussels, Belgium.

Koys, D. (2005). The validity of the Graduate Management Admissions Test® for non-U.S. students. *Journal of Education for Business, 80*(4), 236-239.

Talento-Miller, E. (2005, October). *GMAT® validity for six European MBA programs.* McLean, VA: GMAC®.

Talento-Miller, E., & Rudner, L. M. (2005, October). *GMAT® validity study summary report for 1997 to 2004.* McLean, VA: GMAC®.

Wilson, K. M. (1985). *Factors affecting GMAT® predictive validity for foreign MBA students: An exploratory study.* Princeton, NJ: Educational Testing Service.

