J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 5, Number 4 · December 2006

Differential Item Functioning of GRE Mathematics Items Across Computerized and Paper-and-Pencil Testing Media

Lixiong Gu, Samuel Drake & Edward W. Wolfe

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

JTLA

Volume 5, Number 4

Differential Item Functioning of GRE Mathematics Items Across Computerized and Paper-and-Pencil Testing Media

Lixiong Gu, Samuel Drake & Edward W. Wolfe

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Kevon R. Tucker-Seeley
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Gu, L., Drake, S., & Wolfe, E. W. (2006). Differential item functioning of GRE mathematics items across computerized and paper-and-pencil testing media. *Journal of Technology, Learning, and Assessment, 5*(4). Retrieved [date] from http://www.jtla.org.

J·T·L·A

Abstract:

This study seeks to determine whether item features are related to observed differences in item difficulty (DIF) between computer- and paper-based test delivery media. Examinees responded to 60 quantitative items similar to those found on the GRE general test in either a computer-based or paper-based medium. Thirty-eight percent of the items were flagged for cross-medium DIF, and *post hoc* content analyses were performed focusing on page formatting, mathematical notation, and mathematical content of the items. Although findings suggest that differences in page formatting and response processes across the delivery media contribute little to the observed cross-medium DIF, differences in the mathematical notation contained in the item text as well as differences in the mathematical content of the items provided the strongest apparent relationships with cross-medium DIF.

J·T·L·A

Differential Item Functioning of GRE Mathematics Items Across Computerized and Paper-and-Pencil Testing Media

Lixiong Gu
Samuel Drake
Michigan State University
Edward W. Wolfe
Virginia Tech

Introduction

Computers have become a common tool for both instruction and assessment in educational settings in the United States. And, as is true when any new technology is adopted, efforts are being made to demonstrate that the use of computers in these settings does not disadvantage subgroups of students, particularly in the area of large-scale, high-stakes testing. To date, the research concerning the comparability of scores from computer-based versus paper-based multiple-choice tests has produced mixed results (Bodmann & Robinson, 2004; Choi & Tinkler, 2002; Mason, Patry, & Berstein, 2001). Generally, the differences between average test scores obtained from these two testing media have been small. In addition, there is little evidence that ethnic minorities or females are disadvantaged when they take a computer-based test (Gallagher, Bridgeman, & Cahalan, 2002). However, some evidence suggests that examinees who experience computer anxiety or who have limited computer experience may be slightly disadvantaged when they take computer-based tests, particularly if the examinee must type a response to the test items (Lankford, Bell, & Elias, 1994).

One area that has received little attention to date concerns the degree to which individual test items maintain their comparability between a paper-based and a computer-based medium. A considerable amount of research has been conducted concerning user interface features that facilitate computer-based instruction (for example, Fulcher, 2003; Halima, 2002), but only limited work has been done to determine how key facets of the user interface influence the responses that examinees make to individual test items. Hence, the problem addressed by this paper concerns our lack of understanding of how and whether item features lead to cross-medium performance differences on mathematics multiple-choice test items. By better understanding how item-level features influence the comparability of student performance between computer-based and paper-based media, test developers will be better able to develop test items that allow examinees to demonstrate maximal performance in a computer-based testing medium.

Theoretical Background

The literature concerning the uses of computers in educational settings and in educational testing has revealed several important issues about the relationship between computer experience, computer attitudes, and computer skill. First, despite the widespread use of the Internet and e-mail, there are still inequities in the degree to which some students have access to and familiarity with computers. In the U.S., minorities and females, especially in rural areas, have restricted access (Davies, Klawe, Nhus, Ng, & Sullivan, 2000; Magoun, Eaton, & Owens, 2002; Miller & Varma, 1994). In addition, males tend to dominate computer use in schools and they are also more likely to have access to a computer at home. In contrast, females use the computer less intensively, report a lower level of familiarity with computer applications and have more negative beliefs about computers, in general (Grignon, 1993; van Braak & Kavadias, 2005). Second, these inequities in computer access and familiarity may lead to higher levels of anxiety toward computer-based tasks for disadvantage groups (Colley & Comber, 2003; Massoud, 1992; Shashaani, 1997). As a result, minorities and females not only experience higher levels of computer anxiety but also lower levels of confidence for computer-related tasks (Mitra, Lenzmeier, Steffensmeier, Avon, Qu, & Hazen, 2001). If computer access were equal among all students, this might not be a problem: Research has shown that between group differences in anxiety levels may be diminished when computer experience is held constant.

With respect to the influence of computers on testing in general, there seem to be four important conclusions: First, computer-based tests are more difficult, on average, than conventional tests – although effect sizes indicate that only small differences exist between these two testing media (Ford, Romeo, & Stuckless, 1996; Gallagher, et al., 2002; Mazzeo & Harvey, 1988; Mead & Drasgow, 1993; Russell, 1999). Interestingly, students tend to believe that they will receive higher scores on computer-based tests – a misperception that may drive some students to select a testing medium on which they will receive lower scores (Russell, 1999). Second, even though average differences between test scores from computer-based versus conventional tests are not large, the impact may be great for a small portion of the examinee population (Wise & Plake, 1989). Third, evidence suggests that computer-based testing invokes different cognitive and affective responses on the part of examinees than does conventional testing (Murphy, Long, Holleran, & Esterly, 2000; Lankford, et. al, 1994) and that, unfortunately, affective responses (i.e., computer anxiety and proficiencies, levels of computer experience) are correlated with test scores on computer-based tests at non-trivial levels (Marcoulides, 1988). Fourth, fortunately, training examinees to use the computer interface reduces the negative impact of these variables on computer-based test scores (Johnson & White, 1980; Powers & O'Neill, 1993).

Previous comparability studies have focused mostly on the mode difference at the test level, whereas little work has addressed the direct comparability of items that appear on text versus a computer monitor (Pommerich, 2004). Some insight may be borrowed from human engineering and ergonomics research, which addresses the perceptual and cognitive factors that contribute to differences in reading and problem solving between computer and paper-and-pencil (Castelhano & Muter, 2001; Dillon, 1992; Muter, 1996; Muter & Maurutto, 1991). For example, Muter (1996) identified several differences between text as it appears on a computer monitor versus on paper (i.e., angle of the reading material, shape, actual size, spacing of the characters, line space, margin, polarity) and indicated that these differences influence the processing of text from these two media. Other researchers (Dillon, 1992; Schwarz, Beldie, & Pastoor, 1983) have identified additional variables that may influence the text processing-by-medium interaction – contrast ratio between characters and background, image polarity, text scrolling versus paging, visual navigation, and visual search patterns. Many ergonomic factors have changed rapidly since the 1980s and 1990s when much of the pioneering research took place. However, the current belief is that even with high-quality monitors and dark characters appearing on a light background, reading speed and comprehension is, at best, equivalent for the computer and paper media (Choi & Tinkler, 2002; Muter & Maurutto, 1991).

The study described in this article seeks to identify potential reasons for differential cross-medium item difficulty in mathematics items by performing differential item functioning (DIF) analyses, followed by a detailed content analysis of items similar to those found on the General Test of the Graduate Record Examination™ (GRE™).

Method

Examinees

Data were collected from 165 first-year graduate student volunteers at a large, Research I university in the Midwestern United States. The sample's characteristics are similar to those of graduate student populations at similar institutions. Specifically, the sample contained slightly more males than females (55% versus 45%), a wide mix of academic majors (about 30% were from the Natural Sciences, 30% from Engineering, 15% from the Social Sciences, and the remainder spread across a variety of majors), and a majority of whites and Asians (85% of the sample). In addition, the examinees reported being fairly competent, comfortable, and experienced computer users (i.e., 99% reported using a computer every day, 97% reported being comfortable or very comfortable using a computer, and 92% rated their computer skills as being good or excellent). Finally, nearly all of the examinees indicated that they were comfortable taking computer-based tests (99%) and had previously taken at least one computerized test (89%). This is a very important issue concerning this sample. It is range-restricted toward computer-literate examinees, a fact that should minimize any observed differences between performance on a computer-based and a paper-based test.

Examination

The instrument was created by GRE™ test development staff, who assembled three forms of 20 multiple-choice mathematics items taken from POWERPREP™, which is computer software designed for GRE preparation (ETS®, 1999)[1]. Items were classified into isomorphic types (i.e., items that require similar knowledge and processing), and forms were constructed to be parallel according to the isomorphic structure. Specifically, a trio of items of a common isomorphic type was placed into each of the 20 item positions across the three test forms. Of the 20 items in each form, the first 12 items were classified as being *Quantitative Comparison* items by GRE test developers and the remaining items were classified as being *Problem Solving* items. All items were translated from a computer-based presentation to a paper-based presentation by GRE test developers with the intention of making the two item presentation formats as similar as possible with the exception of text-wrapping (which was allowed to wrap according to column width on the paper-and-pencil test forms). On both administration formats, response options for each item appeared indented and below the item text. For the computer-based test forms, examinees clicked on a bubble using the computer's mouse. For the paper-based test forms, examinees filled in a bubble on a separate answer sheet.

Psychometric analysis of the computer-based and paper-based test data indicated that measures from both forms exhibited comparable – although less-than-ideal – levels of reliability, with the reliability coefficients[2] for the paper-based and the computer-based versions averaging 0.65 and 0.62, respectively. Similarly, average item score-total score correlations for the items on the paper-based and computer-based forms were of similar magnitudes: 0.45 for the paper-based form and 0.44 for the computer-based form. Both instruments exhibited strong primary dimensions in exploratory factor analyses, but that factor accounted for considerably more variance on the computer-based test (56%) than on the paper-based test (43%).

Data Collection

Examinees were recruited by posting flyers on campus offering $50 to participate in the two-hour research session. Examinees were tested in groups of approximately 20 at various times over the course of a semester. The computer-based tests were administered using personal computers in a campus computer lab. Those computers were equipped with central processing units with speeds of at least 500Hz and Internet bandwidths of 10Mb or more. All of the 17-inch color monitors had screen resolutions adjusted to 640x480, which resembled the standard testing settings for the computerized GRE test. The computer-based test was presented in Microsoft Internet Explorer 5.0 and was delivered by XML (Extensible Markup Language) as a means of representing items and models via the Internet. Mathematical expressions in the test were represented with *WebEq*, a free plug-in that enables the Internet browser to display mathematical equations. Examinees recorded their answers to the computer-based forms by clicking an on-screen bubble that appeared next to each answer option.

Each form of the paper-and-pencil test was presented on eight pages of standard, letter-size paper. Each page contained two to four items displayed in 12-point Times New Roman font. For Quantitative Comparison items, the response options appeared at the top of each page for all items on that page. Options for the eight Problem Solving items appeared immediately below each item. Examinees recorded their answers to the paper-based forms on bubble sheets similar to those used in paper-and-pencil GRE test.

During the testing session, examinees completed a demographic questionnaire using the computer and then responded to two of the test forms (one on computer and one using paper-and-pencil), randomly chosen from the three parallel forms. Examinees were required to finish each form within 30 minutes, regardless of the medium type. In addition, medium sequence and test form were counter-balanced to compensate for fatigue and sequence effects.

Analysis

Differential Item Functioning Detection

Differential item functioning (DIF) refers to a class of statistical procedures available to data analysts that allow one to identify whether individual test items exhibit differential levels of difficulty for different demographic groups of examinees who are matched on performance on the test (Camilli & Shepard, 1994; Clauser & Mazor, 1998). We adapted that methodology for the purpose of identifying whether groups of examinees who are matched on test performance and who responded to test items that appeared in a different administration medium (paper versus computer) exhibited differential performance across those administration media. That is, DIF procedures allowed us to identify whether individual test items are more or less difficult for examinees when the items appear in a computer versus a paper medium. Details of the DIF procedures that we utilized are presented in Appendix A. The important point here is that, based on these analyses, we were able to identify items that were more difficult when they appeared on a computer screen and items that were more difficult when they appeared on paper.

Content Analysis

Content analyses were performed on the items flagged for DIF to determine whether some item types were flagged more frequently than were other item types. These content analyses focused on four general features of the test items: a) *verbatim page layout*; b) *mathematical notation*; c) *GRE item classifications*; and d) *mathematical content*. First, we examined both versions of each item to determine whether the fonts, page layouts, and other literal features of the items were identical on the computer-based and paper-based tests (*verbatim page layout*). Specifically, we examined the font size, font spacing, text wrapping, and placement of the answer choices. Second, we examined each item to determine whether the item text contained mathematical nomenclature that might be perceived differently in a paper-based versus computer-based medium (*mathematical notation*). Specifically, we flagged items if they contained any of the following features: a) equations or inequalities; b) variables; c) mathematical operations; or d) were text-based (i.e., required more than a minimal amount of reading). Third, we classified the items into the two item types recognized on the GRE: Quantitative Comparison (QC) and Problem Solving (PS) *(GRE Item classifications)*. Quantitative comparison items are items in which examinees compare the relative sizes of two quantities. Problem Solving items are standard multiple-choice exercises. Fourth we examined the *mathematical content* of the items. The content was categorized as a) arithmetic; b) choose numbers; c) geometric; or d) solve equations.

Results

DIF Outcomes

On average, there was no difference in raw score performance between computer-based (CB) and paper-and-pencil (PP) versions of the tests. Table 1 displays the descriptive statistics for examinee performance (in the raw score metric) and item difficulty (in the IRT metric). The CB version of Form 1 was slightly easier than the PP version. The opposite was true for Forms 2 and 3. However, these differences are small.

Table 1: Raw Score and Item Difficulty Descriptive Statistics

	Form 1		Form 2		Form 3	
	CB	PP	CB	PP	CB	PP
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Raw Score	15.39 (3.28)	14.84 (3.96)	15.25 (3.38)	15.64 (3.70)	15.13 (4.16)	15.94 (3.88)
Item Difficulty	–0.05 (1.45)	0.05 (1.25)	0.10 (0.99)	0.02 (1.08)	–0.04 (1.24)	–0.07 (1.09)
N	54	55	55	55	55	54

A total of 23 items were flagged for cross-medium DIF (38%), of which 11 were more difficult in the paper-and-pencil format, and 12 were more difficult in the computer-based medium. Table 3 (next page) presents descriptive statistics for the item difficulties and the item-total score correlations for each version of each DIF flag classification. Items flagged for DIF were more difficult when they appeared in the medium within which they were flagged (i.e., paper versions of the items flagged for PP DIF have a higher mean difficulty than the computer versions of these items, 0.47 versus –0.25, respectively). In addition, the difference between the means of the two media were nearly equal for the items that were not flagged for DIF as shown in the rightmost column of the first row of data in Table 2. Finally, there were only small differences in the average item-total score correlations between administration media and across DIF flag status.

Table 2: Item Index Means by DIF Flags

		Flagged For		
Index	**Medium**	**PP**	**CB**	**Neither**
Difficulty	**Computer**	–0.25	0.31	–0.03
	Paper	0.47	–0.42	0.00
$r_{\text{item-total}}$	**Computer**	0.39	0.43	0.43
	Paper	0.47	0.47	0.37

Note: Each cell displays the mean of the index.

The three forms of the tests were created to be content-parallel. That is, the three forms constitute 20 sets of items, each set focusing on a common skill and knowledge set. The second and third columns of Table 3 show the average (and standard deviation of the) difficulty of the three items in each trio when those items were administered via paper-and-pencil and when they were administered via computer. The average difficulties are highly correlated between administration format (r=.97, r_{adjusted}=.97). The average absolute difference in item difficulties between administration formats is also fairly small ($M_{|\text{difference}|}$=0.17). However, the standard deviation of the item difficulties within those sets for each administrative format is fairly large, relative to the absolute difference of difficulties between administrative formats (i.e., the average standard deviation of item difficulties within both formats equals 0.40, compared to the average difference of item difficulties between formats of 0.17).

Table 3: Item Set Difficulty Estimates and DIF Status

	Mean$_{CB}$	Mean$_{PP}$			
Item Set	**(SD$_{CB}$)**	**(SD$_{PP}$)**	**Form 1**	**Form 2**	**Form 3**
1	–1.66 (0.64)	–1.38 (0.84)	—	PP	—
2	0.05 (0.58)	–0.55 (0.77)	—	CB	CB
3	–0.48 (0.15)	–0.44 (0.34)	—	CB	—
4	–0.75 (0.70)	–1.02 (0.69)	—	—	—
5	–0.27 (0.48)	–0.39 (0.40)	—	—	CB
6	–0.41 (0.57)	–0.07 (0.16)	PP	—	PP
7	0.65 (0.03)	0.78 (0.07)	—	—	—
8	0.30 (0.14)	0.32 (0.21)	—	—	—
9	0.66 (0.43)	1.05 (0.57)	PP	CB	PP
10	–0.07 (0.14)	–0.30 (0.47)	—	CB	—
11	0.70 (0.31)	0.87 (0.22)	PP	—	—
12	2.69 (0.46)	2.60 (0.50)	—	PP	CB
13	–1.26 (1.00)	–1.25 (0.21)	PP	CB	—
14	–0.62 (0.42)	–0.66 (0.53)	—	—	—
15	–0.39 (0.29)	–0.31 (0.17)	—	—	—
16	0.14 (0.55)	0.05 (0.41)	—	CB	PP
17	–0.22 (0.39)	–0.05 (0.48)	CB	PP	PP
18	–0.21 (0.41)	–0.13 (0.32)	—	—	CB
19	–0.28 (0.14)	–0.41 (0.46)	CB	—	—
20	1.41 (0.24)	1.29 (0.18)	—	—	—

Note: Mean$_{CB}$ and SD$_{CB}$ indicate the mean and standard deviation of the item difficulty estimates in the computer-based medium. PP indicates paper-and-pencil medium. Within each medium, item difficulties were scaled to have a mean of zero and standard deviation of 1.00. The rightmost three columns indicate whether a particular item in the item set was flagged as being more difficult in the computer-based or paper-and-pencil medium. — indicates that the item was not flagged for DIF.

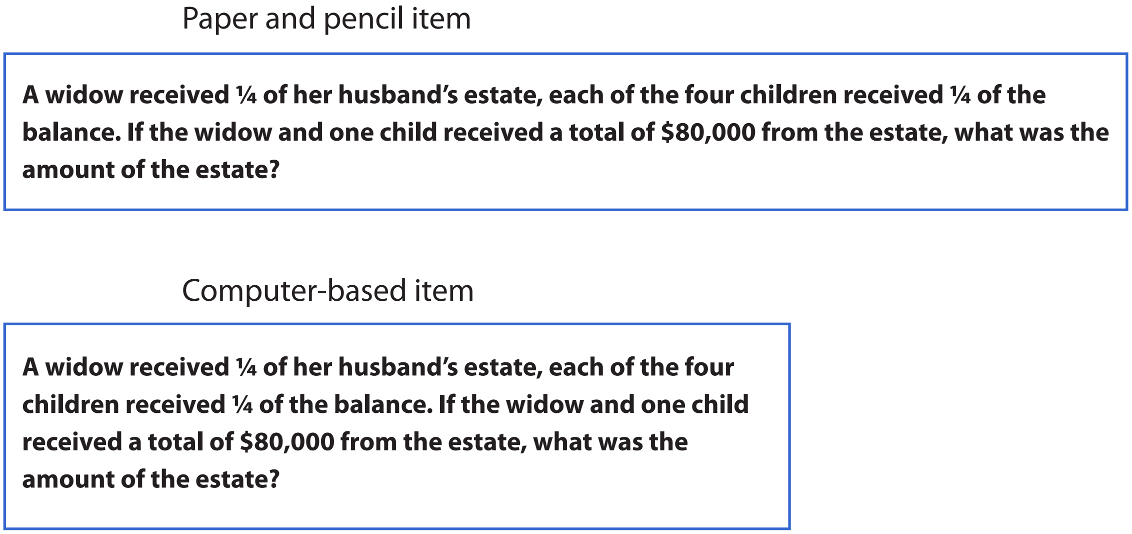
Another important feature of the figures reported in Table 3 concerns the three rightmost columns of the table. These columns indicate which items within an item set were flagged for being more difficult in one medium versus the other. For example, the item in Item Set 1 that appeared on Form 2 was flagged as being more difficult when it appeared in paper-and-pencil than when it appeared on the computer screen. First, note that there are several item sets (six of the 20, to be exact) in which no items were flagged for DIF (i.e., Item Set 4). In general, the differences between the mean difficulty estimates for each administration medium are small within these sets. Second, note that there are several item sets for which there is evidence of a consistent trend in item difficulty between the two administration media. Specifically, 9 of the 20 item sets contained a single item or multiple items flagged for DIF in a single medium. In all but two of these item sets (Item Sets 3 and 18), the mean difference of item difficulties between sets is consistent with the DIF flags. Third, there are also several item sets for which there are inconsistent trends with respect to DIF flags. In addition to the two sets for which the mean item difficulties are inconsistent with the single flagged item, there are five item sets within which multiple, inconsistent DIF flags occurred (i.e., at least one item was flagged as being more difficult in the computer-based medium and at least one other item was flagged as being more difficult in the paper-based medium).

Content Analysis

Verbatim Page Layout Formatting

Examination of the appearance of the items in the computer-based and paper-based media indicated that 21 of the 60 items (35%) exhibited formatting differences between the two media. The most common difference was that of text wrapping. More characters could fit on a paper-and-pencil line, thus the computer-based items tended to have more text lines per item than the paper-and-pencil items. Figure 1 displays an example of how an item may have appeared in paper and pencil format and the same item in computer-based format.

Figure 1: Text Wrapping Examples



Paper and pencil item

A widow received ¼ of her husband's estate, each of the four children received ¼ of the
balance. If the widow and one child received a total of $80,000 from the estate, what was the
amount of the estate?

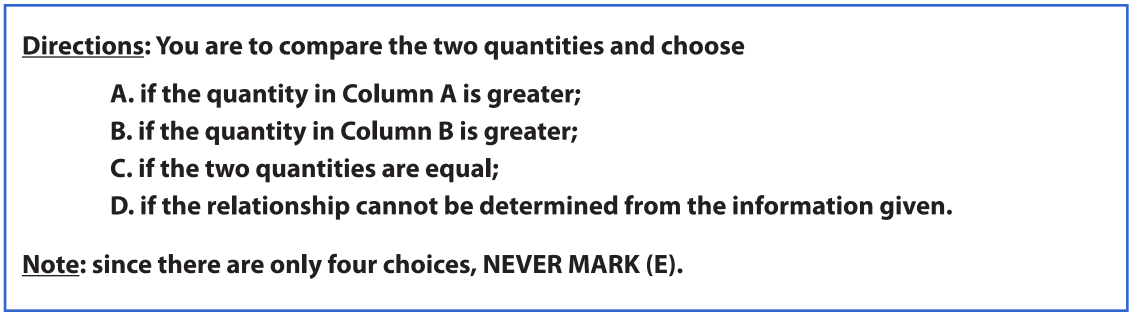
Computer-based item

A widow received ¼ of her husband's estate, each of the four
children received ¼ of the balance. If the widow and one child
received a total of $80,000 from the estate, what was the
amount of the estate?

There was another pervasive difference in the formatting of the quantitative comparison items between media. Specifically, on the paper-based tests, the answer choices and their meanings are placed at the top of each page, and multiple items may appear on a page. However, on the computer-based test, the answer choices immediately follow each question, and only one item appears on each screen. Also, for the problem solving items, examinees choose between five answer choices labeled A, B, C, D, and E on the paper-and-pencil version. On the other hand, the five answer choices appear as a bubble (formatted as 0), on the computer-based version, requiring examinees to click on the bubble to select that option.

Figure 2: Answer Choice Examples

Paper and pencil layout

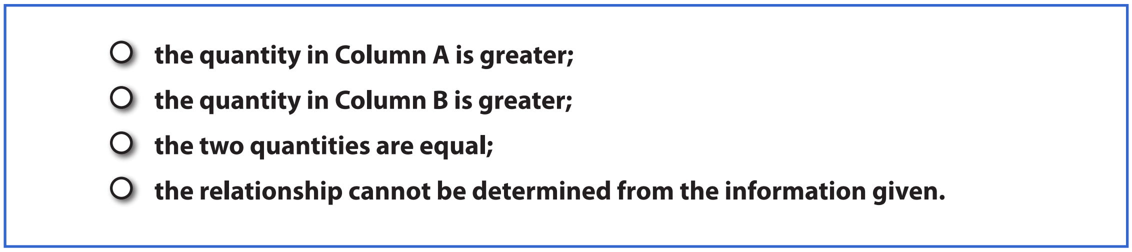


<u>Directions</u>: You are to compare the two quantities and choose

A. if the quantity in Column A is greater;
B. if the quantity in Column B is greater;
C. if the two quantities are equal;
D. if the relationship cannot be determined from the information given.

<u>Note</u>: since there are only four choices, NEVER MARK (E).

Computer layout



○ **the quantity in Column A is greater;**
○ **the quantity in Column B is greater;**
○ **the two quantities are equal;**
○ **the relationship cannot be determined from the information given.**

Table 4 summarizes our comparison of the layout of the paper-and-pencil versus the computer-based versions of the test items. Overall, 43% of the flagged items exhibited formatting differences, and 30% of the non-flagged items exhibited formatting differences. Of the 11 items flagged as being more difficult on paper and pencil, 5 (45%) exhibited minor differences in the item layout. Of the 12 items that were flagged as being difficult on computer, 5 (42%) exhibited minor differences in item layout.

Table 4: Format Differences by DIF Flags

	Flagged For			
Format	**PP**	**CB**	**Neither**	**Total**
Identical	6	7	26	**39**
Different	5	5	11	**21**
Total	**11**	**12**	**37**	**20**

Note: Each cell displays the number of items falling into that category.

Mathematical Notation of the Test Items

Table 5 (next page) summarizes our analyses of the mathematical notation of the items as they relate to DIF. Overall, 45% of the items involved solving an equation or an inequality, 60% contained notation relating to variables, 80% contained operations notation, and 50% of the items were text-based. A larger percentage of the items flagged for PP DIF contained equations than did the items flagged for CB DIF (73% versus 27%). The same was true for items containing variable designations (82% versus 42%). There were no apparent trends in the relationship between DIF flag rates and the existence of mathematical operators or text in the items.

Table 5: Mathematical Notation Differences by DIF Flags

		Flag			
Feature	**Present**	**PP**	**CB**	**Neither**	**Total**
Equality	Yes	8	3	16	27
	No	3	9	21	33
	Total	**11**	**12**	**37**	**60**
Variable	Yes	9	5	22	36
	No	2	7	15	24
	Total	**11**	**12**	**37**	**60**
Operations	Yes	10	9	29	48
	No	1	3	8	12
	Total	**11**	**12**	**37**	**60**
Text-based	Yes	4	6	20	30
	No	7	6	17	30
	Total	**11**	**12**	**37**	**60**

Note: Each cell displays the number of items falling into that category.

GRE Item Content Classifications

Table 6 summarizes the relationship between DIF flags and the GRE item content classifications. 60% of the items were classified as Quantitative Comparisons, and 40% were classified as Problem Solving items. The empirical percentages of items flagged for DIF differed only slightly from these marginal percentages (64% and 36% for PP versus 58% and 42% for CB).

Table 6: GRE Item Classification by DIF

GRE Item Type	PP	CB	Neither	Total
Quantitative Comparison	7	7	22	**36**
Problem Solving	4	5	15	**24**
Total	**11**	**12**	**37**	**60**

Note: Each cell displays the number of items falling into that category.

Mathematical Content of the Items

Table 7 (next page) summarizes our analyses of the mathematical content of the items as it relates to cross-medium DIF. In all categories, 20% of the items contained the type of mathematical content in question (i.e., arithmetic, solving equations, or choosing numbers). 30% of the items (18 of 60) contained arithmetic computations, and 50% of these items required knowledge of decimals, 33% required operations with fractions, and about 17% required knowledge of square roots. The figures indicate that a higher proportion of items requiring arithmetic were flagged for DIF, with a greater proportion of those flagged items being flagged for being more difficult in the computer-based medium. Specifically, 58% of the items flagged for CB DIF contained arithmetic and 36% of the items flagged for PP DIF, compared to only 19% of the non-flagged items containing arithmetic. There were only small differences between the marginal percentages of items classified as relying on equation solution or number choice and the conditional percentages for DIF flag status.

Table 7: Mathematics Content by DIF

Feature	Present	PP	CB	Neither	Total
Arithmetic	Yes	4	7	7	12
	No	7	5	30	48
	Total	**11**	**12**	**37**	**60**
Solve Equation	Yes	4	2	6	12
	No	7	10	21	48
	Total	**11**	**12**	**37**	**60**
Choose Number	Yes	5	3	4	12
	No	6	9	33	48
	Total	**11**	**12**	**37**	**60**

Note: Each cell displays the number of items falling into that category.

Discussion

In this exploratory study, we attempted to identify explanations for cross-medium differential item functioning in GRE mathematics items. Because our sample of items was small and because of the *post hoc* nature of our analyses, we formulate our discussion of the results as observations that can be used to generate tentative hypotheses for future exploration. We present five such observations.

First, it is somewhat remarkable that cross-medium DIF exists in this sample. The sample is homogeneous in terms of computer experience – it contains only graduate students at a Research I university with good representation of students in technical fields like Engineering and Natural Sciences, and most examinees had considerable experience and comfort in both computer use and computerized testing. As a result, we would expect this sample to underestimate potential differences in cross-medium performance. The fact that 38% of the items exhibited cross-medium DIF suggests that the medium in which items appear may have an important impact on the manifest difficulty of that test item. However, consistent with other studies of cross-medium performance on multiple-choice tests (Gallagher et al., 2002; Mead & Drasgow, 1993), we did not find a systematic trend toward items being more difficult in one medium versus another medium – some items are more difficult when they appear on paper, and some items are more difficult when they appear on a computer monitor.

Second, with respect to the formatting of the two instruments, we found them to be highly similar. This is not surprising, given that the test developers took great care to make the instruments as identical as possible. The two most apparent differences in formatting related to word-wrapping (due to different line lengths on the monitor versus printed page) and to position and appearance of the responses for these multiple-choice questions (i.e., presentation of the Quantitative Comparison options only at the top of each page on the paper-and-pencil test versus on each screen for the computer-based test and filling in lettered bubbles for the paper-and-pencil test versus clicking a non-lettered bubble on the computer-based test). However, we saw no apparent relationship between DIF flag rates and this level of item formatting. Hence, we conclude that difference in surface-level appearance and response processes across the delivery media is an unlikely explanation for the observed DIF.

Third, there seem to be few consistent patterns in the DIF flags of items designed to be similar to one another. We observed considerable variability with respect to flag rates and item difficulties within item sets that were constructed to be parallel in terms of content. Because these items were content equivalent, they would also tend to be coded into similar content analysis categories within an item set. Hence, there seems to be little variability that can be explained in terms of general item content. In fact, we observed no apparent relationships between the GRE item classifications and DIF flag rates. However, we should point out that there is considerable variability in the required processing and mathematical content within the GRE classifications. On the other hand, our own more detailed categorization of the mathematical content suggests that items targeting arithmetic skills are more likely to be flagged as being more difficult in the computer-based medium whereas items requiring the examinee to choose a number are *less* likely to be flagged as being more difficult in the computer-based medium. We speculate that the presentation of items in different delivery media invokes differences in the cognitive approaches taken when examinees respond to test items.

Fourth, there seem to be differences in DIF flag rates attributable to the mathematical notation contained in a test item. Specifically, items containing equalities/inequalities and items containing variables were more likely to be flagged for being more difficult on paper. On the other hand, there were no apparent trends relating to DIF flag rates of the existence of operators or text-based item notation. Hence, we can speculate that differences in the appearance of mathematical notation between the computer screen and paper is a reasonable explanation for the observed cross-medium DIF.

Finally, we should point out that our analyses may not have sufficiently captured the complexity of these data. We did not investigate interactions between the content categories that we observed due to the small number of items we considered. For example, it may be that item features such as verbatim formatting and mathematical content are related and that these features produce a cumulative effect on cross-medium DIF. As a result, we believe that additional studies on large item pools are warranted to determine whether the observed DIF flag rates can be replicated in other samples of examinees and whether item features interact to produce cross-medium DIF.

Endnotes

1. These data were collected as part of a larger project studying the influence of isomorphic items on examinees' perceptions of test items (Morely, Bridgeman, & Lawless, 2003). In exchange for providing facilities and staff for that project, those authors agreed to allow for the collection of the additional data presented in this manuscript.
2. Reliability indices were computed by *Winsteps* (Linacre, 2002) as 1 – [MSE / V(Estimates)].

References

Bodmann, S. M. & Robinson, D. H. (2004). *Speed and performance differences among computer-based and paper-pencil tests*. Baywood Publishing Co., Inc.

Camilli, G. & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Castelhano, M. S. & Muter, P. (2001). Optimizing the reading of electronic text using rapid serial visual presentation. *Behaviour & Information Technology*, *20*, 237–247.

Choi, S. W. & Tinkler, T. (2002). *Evaluating comparability of paper and computer-based assessment in a K–12 setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Clauser, B. E. & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*, 31–44.

Colley, A. & Comber, C. (2003). Age and gender differences in computer use and attitudes among secondary school students: What has changed? E*ducational Research, 45*(2), 155–165.

Davies, A., Klawe, M., Nhus, C., Ng, N., & Sullivan, H. (2000). *Gender issues in computer science education: Beyond description of the problems*. Washington, DC: National Center for Improving Science Education.

Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics, 35*, 1297–1326.

Draba, R. E. (1977). *The identification and interpretation of item bias* (Research Memorandum No. 26). Chicago, IL: University of Chicago.

ETS. (1999). GRE General Test Powerprep® Software (Version 2.0). Princeton, NJ: Educational Testing Service.

Ford, B. D., Romeo, V., & Stuckless, N. (1996). The effects of computer versus paper-and-pencil administration on measures of anger and revenge with an inmate population. *Computers in Human Behavior, 12*, 159–166.

Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing, 20*(4). 384–408.

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

Grignon, J. R. (1993). Computer experience of Menominee Indian students: Gender differences in coursework and use of software. *Journal of American Indian Education, 32*, 1–15.

Halima, H. M. (2002). *Cognitive user profile and its involvement into adaptive interface*. Paper presented at E-Learn 2002 World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education, Montreal, Quebec, Canada.

Johnson, D. F. & White, C. B. (1980). Effects of training on computerized test performance in elderly. *Journal of Applied Psychology, 65*, 357–358.

Lankford, J. S., Bell, R. W., & Elias, J. W. (1994). Computerized versus standard personality measures: Equivalency, computer anxiety, and gender differences. *Computers in Human Behavior, 10*, 497–510.

Linacre, J. M. (2002). A user's guide to WINSTEPS/MINISTEP Rasch-model computer programs (Version 3.36). Chicago, IL: MESA Press.

Loyd, B. H. & Gressard, C. P. (1986). Gender and amount of computer experience of teachers in staff development programs: Effects on computer attitudes and perceptions of usefulness of computers. *AEDS Journal, 19*, 302–311.

Magoun, D., Eaton, V., & Owens, C. (2002). *IT and the attitudes of middle school girls: A follow-up study*. Paper presented at the NECC 2002: National Educational Computing Conference Proceedings, San Antonio, TX.

Marcoulides, G. A. (1988). The relationship between computer anxiety and computer achievement. *Journal of Educational Computing Research, 4*, 151–158.

Mason, B. J., Patry, M., & Berstein, D. J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. *Journal of Educational Computing Research, 24*, 29–39.

Massoud, S. L. (1992). Computer attitudes and computer knowledge of adult students. *Journal of Educational Computing Research,* 7, 269–291.

Mazzeo, J. & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional versions of educational and psychological tests: A review of the literature* (No. CBR 87–8). Princeton, NJ: Educational Testing Service.

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449–458.

Miller, F. & Varma, N. (1994). The effects of psychosocial factors on Indian children's attitudes toward computers. *Journal of Educational Computing Research, 10*, 223–238.

Mitra, A., Lenzmeier, S., Steffensmeier, T., Avon, R., Qu, N. & Hazen, M. (2001). Gender and computer use in an academic institution: Report from a longitudinal study. *Journal of Educational Computing Research, 23*, 67–84.

Morely, M. E., Bridgeman, B., & Lawless, R. R. (2003). *Impact on the Use of Item Variants on Math Test Performance*. Paper presented at the National Council on Measurement in Education, Chicago, IL.

Muter, P. (1996). Interface design and optimization of reading of continuous text. In H. van Oostendorp & S. de Mul (Eds.), *Cognitive aspects of electronic text processing* (pp. 161–180). Norwood, NJ: Ablex.

Muter, P. & Maurutto, P. (1991). Reading and skimming from computer screens and books: The paperless office revisited? *Behaviour and Information Technology, 10*, 257–266.

Powers, D. E. & O'Neill, K. (1993). Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. *Educational Assessment, 1*, 153–173.

Pommerich, M. (2004). Developing computerized versions of paper tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2*(6), 1–44.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197–207.

Ronau, R. N. & Battista, M. T. (1988). Microcomputer versus paper-and-pencil testing of student errors in ratio and proportions. *Journal of Computers in Mathematics and Science Teaching, 8*, 33–38.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper 20. *Education Policy Analysis Archives,* 7(20). Retrieved from http://epaa.asu.edu/epaa/v7n20.

Schwarz, E., Beldie, I. P., & Pastoor, S. A. (1983). A comparison of paging and scrolling for changing screen contents by inexperienced users. *Human Factors, 25*, 279–282.

Shashaani, L. (1997). Gender differences in computer attitudes and use among college students. *Journal of Educational Computing Research, 16*, 37–51.

Signer, B. R. (1991). CAI and at-risk minority urban high school students. *Journal of Research on Computing in Education, 24*, 189–203.

Temple, L. & Lips, H. M. (1989). Gender differences and similarities in attitudes toward computers. *Computers in Human Behavior, 5*, 215–226.

Van Braak, J. & Kavadias, D. (2005). The Influence of Social-demographic determinants on secondary school children's computer use, experience, beliefs and competence. *Technology, Pedagogy and Education, 14*(1), 43–59.

Wise, S. L. & Plake, B. S. (1989). Research on the effects of administering tests via computers. *Educational measurement: Issues and practice, 8*(3), 5–10.

Wolfe, E. W., Gu, L., & Drake, S. (2003). *Differential item functioning of GRE mathematics items across computerized and paper-and-pencil testing media*. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA.

Wright, B. D. & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, IL: MESA Press.

Appendix A

Differential Item Functioning

Computer-based and paper-based item responses were separately scaled to the Rasch model (Wright & Masters, 1982; Wright & Stone, 1979), and item parameters were estimated separately for each administration medium using *Winsteps* (Linacre, 2002). That is, we generated a separate item difficulty index (δ, expressed on the logit scale – the log of the odds of answering the item correctly versus incorrectly) and standard error for that estimate (SE_δ) for each item in each administration medium (CB versus PP). Raju's Signed Area Index (SAI) (Raju, 1988, 1990) was then computed for each item to determine the magnitude of the shift in item difficulty between the two administration media, $SAI = \delta_{CB} - \delta_{PP}$. Using Draba's criterion (Draba, 1977), items were flagged for exhibiting DIF when the absolute value of the *SAI* was greater than 0.50 logits. Because of the relatively small sample size in our study, we chose not to consider the statistical significance of the SAI, which is computed using the following formula.

$$z_{SAI} = \frac{\delta_{CB} - \delta_{PP}}{\sqrt{SE^2_{\delta_{CB}} + SE^2_{\delta_{PP}}}}$$

Appendix B

Content Analysis Examples

The following were codes used in our analysis.

I. Mathematical Notation

A. *Equality* – items that contain one of the following symbols: $= \quad \leq \quad \geq \quad > \quad <$

Example: Which of the following options is equivalent to $6w + 14 > (–8w) – 56$?

(A) $w < 28$

(B) $w < (–7)$

(C) $w < -5$

(D) $w < 5$

(E) $w > 5$

B. *Variables* – items that contain variable designations (i.e., x)

Example 2: If w exceeds y by 19, then y =\

(A) $x – 19$

(B) $–20$

(C) $–19$

(D) $19x$

(E) $19x + 19$

C. *Operations* – items that contain one or more arithmetic operation symbol (i.e., + × ÷ −)

Example:

If $x+\frac{1}{y}\neq 0$, then witch of the following is equal to the reciprocal of $x+\frac{1}{y}$

(A) $\frac{1}{x}+y$

(B) $\frac{1}{y}-x$

(C) $\frac{y}{xy+1}$

(D) $\frac{x}{xy+1}$

(E) $\frac{1}{x}+\frac{1}{y}$

D. *Text-based* – Word problems or exercises that contain sentences

Example: The ages of three people are such that the age of one person is three times the age of the second person and half the age of a third person. If the sum of their ages is 10, then the age of the younger person is.

(A) 1

(B) 2

(C) 3

(D) 5

(E) 6

II. Mathematical Content

A. *Arithmetic* – items that require computations with real numbers

Example:

Column A	Column B
$\sqrt{2(8)}$	4

B. *Solve Equation* – items that require solving equations or inequalities

Example: If $4x - 2 = 8$ then $8x - 3 =$ _____

(A) 2.5

(B) 5

(C) 10

(D) 17

(E) 20

C. *Choose numbers* – items that require choosing a number and using it to evaluate an expression

Example: x and y are positive integers

$$x < 2$$

$$y > 1$$

Column A	Column B
x	$\frac{y}{2}$

Author Note

This project greatly benefited from the input of our colleagues. Specifically, we want to thank: René Lawless for the incredible support she provided to us in managing this project; Dr. Mary Morley for authoring all of the items that comprised the instrument; Mike Wagner for creating the delivery system used to administer the tests via the Internet; Dr. Mark Reckase and Dr. Terry Ackerman for helpful comments on an early version of this article; and Chris Chan for reviewing a final draft of the manuscript. Finally, we want to point out that this report is based on data that was presented in a paper at professional conferences (see Wolfe, Gu, and Drake, 2003).

Correspondence concerning this article should be addressed to Lixiong Gu, Measurement and Quantitative Methods, Michigan State University, East Lansing, MI 48824; e-mail: gulixion@msu.edu.

Author Biographies

Lixiong Gu is a Ph.D. candidate in Measurement and Quantitative Methods at Michigan State University. His research interests include computer-based testing, item pool design for computerized adaptive tests, and K-12 large scale assessment.

Samuel Drake is a Measurement and Quantitative Methods doctoral students at Michigan State University. He is also a Mathematics Learning Specialist at Michigan State University. His interests include transition from high school to college mathematics and constructing tasks that assess student's understanding of mathematical concepts.

Edward W. Wolfe is an Associate Professor of educational research and evaluation at Virginia Tech. Dr. Wolfe's research focuses on applications of Rasch models to instrument development and the analysis of ratings, influences of technology in testing on examinee mental states, and differential item functioning evoked by test translation.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College