GRADUATE MANAGEMENT ADMISSION COUNCIL

Examining Test Speededness by Native Language

Eileen Talento-Miller, Fanmin Guo, Kyung T. Han

GMAC® Research Reports • RR-12-01 • July31, 2012

Abstract

When power tests include a time limit, it is important to assess the possibility of *speededness* for examinees. Research on differential speededness in the past has included looking at gender and ethnic subgroups in the United States on paper and pencil tests. The needs of a global audience necessitated, and the availability of computer recorded response time data enabled this investigation of differential speededness by native language. Data from a computerized adaptive test administered globally in English were used to compare different measures of time pressure and speed for 15 language groups compared with native-English-speaking examinees. Results indicated that examinees whose native language was Korean had issues with speededness when many of the metrics were considered, but when ability was controlled, there were no relevant differences for any of the languages compared to English. Future research should include independent measures of ability and English proficiency as well as any possible effects of speededness on predictive validity.

Following the definitions of speed and power tests, typical standardized tests do not fall strictly into either category. Generally speaking, the difference in speed and power tests is how scores are determined: number reached or number right, respectively (Lord & Novick, 1968; Rindler, 1979). Though the importance of getting the answers right is emphasized for many tests, time limits are imposed as one of the features ensuring standardized administration, and penalties apply when examinees do not answer all questions, making the test partially speeded (Lord & Novick, 1968; Rindler, 1979). If speed of responding is not part of the construct being measured by the standardized test, it becomes incumbent upon test sponsors to ensure that time limits are used appropriately to preserve construct validity (Bridgeman, 2004; Lu & Sireci, 2007).

Previous research has examined whether subgroups are differentially affected by time limits. Research by Lawrence (1993) suggested differential speededness exists by gender groups and US ethnic subgroups, however, research by Bridgeman (2004) suggested score differences among these groups would remain relatively unchanged with additional time, since examinees benefited from the extra time, regardless of group. Other studies examining primarily race/ethnicity differences in speededness hypothesized that strategies or instructions may have an effect on observed differences. For instance, Dorans, Schmitt, and Bleistein (1988) suggested groups may have differing omit patterns, and Evans and Reilly (1973) suggested some groups may be less likely to guess. Some of the issues regarding the assessment of differential speededness involve the treatment of item difficulty and item order, as well as issues of omits versus items not reached (Dorans, Schmitt, & Bleistein, 1988; Evans & Reilly, 1973; Lawrence, 1993; Peterson, 1993; Rindler, 1973). Previous studies such as these on differential speededness have focused on paper-based administration, which limits the measurement of speededness to observable behaviors such as items left unmarked, whether within or at the end of a section.

Computerized adaptive testing (CAT) has been attractive in part because of measurement efficiency (i.e., improved measurement precision with shorter test length). Like other types of computer-based testing, CAT allows more detailed investigations of how time is spent during test administration. Questions such as, How much time does an examinee have left for the last five questions? and, How many

questions does an examinee have left when there are only two minutes remaining?' can be addressed looking at data from computer administered tests. Some recent studies have examined time pressure on exams, expanding from simple omit and not reached behaviors, to speed of responding at different points during the test (Bridgeman & Cline, 2004; Talento-Miller & Guo, 2009). Notwithstanding the benefits of computer administration in the tracking of behaviors related to speed, there are criticisms of CAT suggesting that the inconsistency in items across examinees may also lead to differences related to time (Chang, 2007; Lu & Sireci, 2007; Guo, 2011; Schmidt, Sass, Sullivan, & Walker, 2010). The possible difference in time required by item difficulty underscores the importance of assessing differences in speed by ability level, such as through the standardization method defined by Dorans, Schmitt, & Bleistein (1988).

Speededness is a potential source of error variance that may become a serious threat to validity, the appropriate interpretation of scores (Lu & Sireci, 2007). A variety of factors may contribute to an individual examinee experiencing time pressure on a test, including personal characteristics, such as risk aversion (Mislevy & Wu, 1996), or demographics such as native language or culture (Emengou & Childs, 2005; Pennock-Román, 1992). For instance, the study by Emengou & Childs (2005) suggested that students taking a Canadian exam in either English or French differed in the number of items answered as well as relative accuracy, suggesting a difference in guessing behavior. The review by Pennock-Román (1992) on Hispanic students suggested speededness may be a factor in assessment for students whose best language is not English, but acknowledged that in some cases variance due to speed may be relevant. The efficacy of test scores varied with level of English proficiency and subject areas (Pennock-Román, 1992). Studies have suggested that there are no significant differences in predictive validity for Hispanic students compared to non-Hispanic white students in the United States (Pennock-Román, 1992; Sireci & Talento-Miller, 2006). A study by Talento-Miller (2008) that examined predictive validity across different citizenship and language groups in non-US schools suggested admission test scores were effective across the groups studied. The study included limited groupings as well as limited schools, however, which make generalizations difficult.

For predicting performance in programs taught in English, administering the test in English is appropriate for interpretations of scores. Beyond the content itself, however, native language may have an effect on the speed of responses to items. Each language varies based on its degree of difference to grammatical conventions and the alphabet used in English. Where possible, an investigation into differential speededness by native language should evaluate many different languages compared to English to sort out the effects and their possible causes. The current study investigates test speededness by native language for a computerized adaptive test used around the world.

Methodology

The Graduate Management Admission Test® (GMAT®) is used globally to aid graduate business programs with admissions decisions. Although the exam is available only in English, as it is intended for programs taught in English, it is taken by more than 250,000 examinees annually and is accepted by more than 5,400 programs around the world. The availability of a large global audience and computer-recorded speed data for each examinee enables research regarding speededness for different language groups.

The GMAT exam consists of three computer administered sections, two of which are adaptive. The verbal and quantitative sections are each allowed a maximum of 75 minutes to answer 41 and 37 questions, respectively. The adaptive algorithm does not allow items to be skipped and scores are adjusted for items not answered when time expires.

Administrations of the GMAT exam in 2009 served as the data for the investigation. Languages with greater than 1,000 cases were included in the comparisons with the focal group of native-English-speaking examinees. Cases were excluded when questionable effort was exhibited, as defined by amount of time spent overall or on early items. For instance, if less than one third of the time allotted was spent in any of the sections, the entire case was removed from the dataset. If extensive time was spent on the first few items (e.g. 20 minutes on one item, more than 40 minutes on the first five items) then the case was

removed, based on the premise that these examinees were intentionally following a differential use of initial time during the test administration in order to memorize items, or try to game the adaptive algorithm. Because examinees can take the GMAT exam more than once in a calendar year, an effort was made to remove repeat cases. Within the dataset, the earliest instance of a case was retained, and subsequent administrations for the same examinee were removed. This means each individual was represented only once in the dataset, but because only one year is examined, it does not necessarily mean the record represents their first time taking the test. Data included latency and accuracy for each item position, as well as total latency and scaled scores by section.

Investigators used several measures that could indicate speededness or time pressure. Average time spent as well as average number of items completed on each section for the group whose native language was English were compared to each of the 15 language groups. Cohen's *d* was used to calculate effect size, using the mean and standard deviation for the English language group. Rule of thumb definitions for speededness were also evaluated, which specify that a test is not speeded if 80% of examinees finish the test and all examinees finish 75% of the test (Evans & Reilly, 1973; Peterson, 1993; Rindler, 1979). In addition, time pressure was evaluated similar to the study by Bridgeman and Cline (2004) by looking at time remaining for the last five questions. Instances of rapid guessing at the end of the section were defined based on procedures described in Talento-Miller and Guo (2009) as consecutive responses with latencies less than 7 or 10 seconds for the quantitative and verbal items, respectively. Specifically, investigators compared the proportion of candidates who had less than two minutes remaining for the last five questions. They also evaluated the difference in proportions of candidates with consecutive rapid guesses starting at each of the last five item positions. Finally, differential speededness was measured using the standardization approach, which calculates the difference in rates of not-reached items for the reference and focal groups by ability level (Dorans, Schmitt, & Bleistein, 1988). For the current study, ability level was defined by scaled score for the section, and the focal group consisted of the examinees whose native language is English.

Results

In addition to native English speakers, there were 15 language groups with more than 1,000 valid cases in the test year data. Table 1 lists the sample sizes for each group, the mean time for each section, the mean number of items completed, and the effect sizes for each when compared with English. Generally, there were fewer differences observed in the quantitative section compared with the verbal section of the test. There were no effect sizes greater than 0.5 in mean time or average number of items completed for the quantitative section. Ten of the 15 languages, however, had an average time greater than the English group on the verbal section, with an effect size greater than 0.5. The largest effects were found for Korean speakers (d = 0.71, n = 4,035) and Japanese speakers (d = 0.70, n = 1,503). By contrast, only one language had more than half a standard deviation difference in average items completed on the verbal section. Compared to English, Korean language examinees answered fewer questions with a very large effect size (d = –4.27, n = 4035).

Table 1. Descriptive Statistics By Language Compared to English					
		Verbal (75 minutes; 41 Items)		Quantitative (75 minutes; 37 Items)	
Language	N	Mean Time in Minutes (Effect Size)	Mean Number Answered (Effect Size)	Mean Time in Minutes (Effect Size)	Mean Number Answered (Effect Size)
English	114,042	67.25; SD = 9.13	40.87; SD = 0.832	71.33; SD = 6.98	36.56; SD = 1.565
Arabic	4,299	70.41 (0.35)	40.67 (-0.24)	71.74 (0.06)	36.30 (-0.17)
Mandarin	26,881	73.24 (0.66)	40.46 (-0.49)	70.60 (-0.10)	36.67 (0.07)
German	3,312	71.90 (0.51)	40.87 (0.00)	73.04 (0.24)	36.72 (0.10)
Spanish	7,557	71.68 (0.49)	40.71 (-0.19)	73.01 (0.24)	36.34 (-0.14)
French	3,925	72.04 (0.52)	40.61 (-0.31)	73.32 (0.28)	36.25 (-0.20)
Guajarati	1,853	71.73 (0.49)	40.85 (-0.02)	72.32 (0.14)	36.67 (0.07)
Hindi	12,464	72.86 (0.61)	40.87 (0.00)	73.26 (0.28)	36.78 (0.14)
Italian	1,100	71.58 (0.47)	40.79 (-0.10)	73.40 (0.30)	36.48 (-0.05)
Japanese	1,504	73.63 (0.70)	40.60 (-0.32)	73.14 (0.26)	36.65 (0.05)
Korean	4,036	73.76 (0.71)	37.32 (-4.27)	72.20 (0.12)	35.98 (-0.37)
Portuguese	1,868	72.37 (0.56)	40.82 (-0.06)	73.43 (0.30)	36.56 (0.00)
Russian	3,400	71.32 (0.45)	40.80 (-0.08)	72.68 (0.19)	36.61 (0.03)
Thai	1,514	73.13 (0.64)	40.71 (-0.19)	73.69 (0.34)	36.37 (-0.12)
Turkish	1,623	71.82 (0.50)	40.66 (-0.25)	71.90 (0.08)	36.57 (0.01)
Vietnamese	1,466	71.84 (0.50)	40.70 (-0.20)	72.68 (0.19)	36.39 (-0.11)

Table 2 shows the results for the percentage of examinees who finished all items and the percentage of examinees who finished at least 75% of items by section and overall. Again, the pattern showed more difficulties encountered in the verbal section, and the group of native Korean speakers often represented outliers from the other data. Using the 80% finish-all-items rule of thumb threshold, both the verbal and quantitative sections would be considered speeded for the native Korean speakers, with completion rates of 47% and 75%, respectively. The only other language falling under 80% was French at 79% in completion in the quantitative section. More than 99% of examinees finished more than 75% of the combined verbal and quantitative items for 14 of the languages. The percentage of Korean examinees that finished at least 75% of the combined two sections was 97%, with values of 88% and 98%, respectively, for the separate verbal and quantitative sections.

Using the time pressure definition of less than two minutes remaining for the last five items, the difference in proportion of examinees by language was examined relative to English. The difference in proportion exceeded 0.1 compared to English for the languages of Korean (0.19), Japanese (0.17), and Thai (0.12) on the verbal section, and Thai (0.10) on the quantitative section. The other time pressure measure—difference in proportion of examinees exhibiting rapid guessing behavior toward the end of the test—showed a very different pattern with no differences greater than 0.1 for any of the languages in either section. These results are illustrated in Figures 1 through 4. As previous research has suggested, cultural differences may affect willingness to guess, however, there may also be notable differences related to ability.

Table 2. Speededness Measures by Language					
		Verbal		Quantitative	
Language	N	% Completing All	% Completing 75%	% Completing All	% Completing 75%
English	114,042	95.4	99.9	85.9	99.2
Arabic	4,299	90.0	99.6	80.6	98.4
Mandarin	26,881	83.4	99.3	87.8	99.6
German	3,312	94.1	99.9	89.2	99.6
Spanish	7,557	91.7	99.6	82.5	98.3
French	3,925	88.7	99.4	78.7	98.2
Guajarati	1,853	94.0	99.7	88.1	99.6
Hindi	12,464	94.7	99.9	90.9	99.8
Italian	1,100	91.8	99.7	83.4	99.0
Japanese	1,504	88.6	98.9	86.8	99.7
Korean	4,036	47.4	87.9	75.2	97.6
Portuguese	1,868	94.0	99.5	86.4	98.8
Russian	3,400	93.2	99.7	87.4	99.3
Thai	1,514	91.1	99.5	80.2	99.1
Turkish	1,623	89.2	99.6	84.6	99.4
Vietnamese	1,466	90.3	99.5	81.5	98.5

Figure 1. Difference in Proportion of Candidates With Time Pressure on Verbal Compared to English

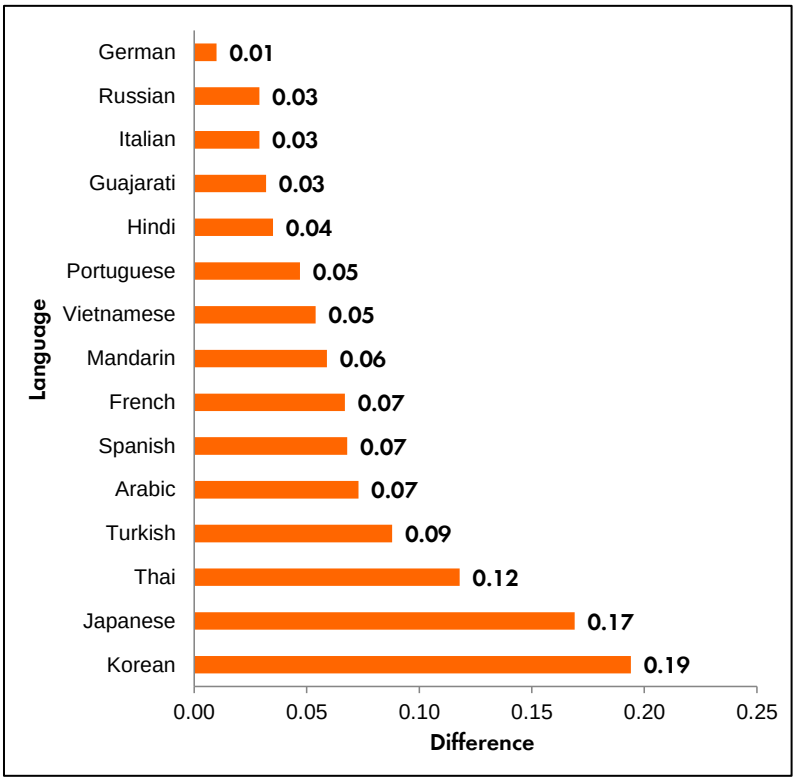


Figure 2. Difference in Proportion of Candidates With Time Pressure on Quant Compared to English

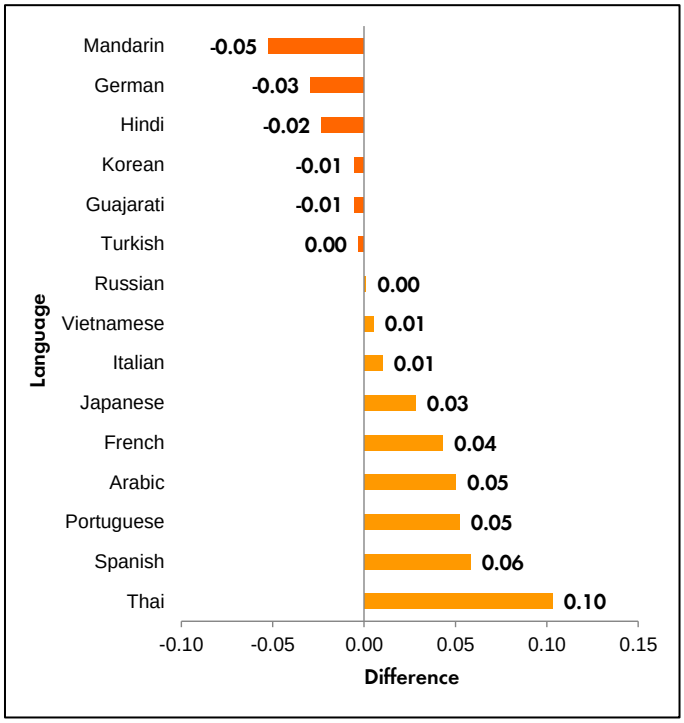


Figure 3. Difference in Proportion of Examinees Guessing on Verbal by Item Position Compared to English

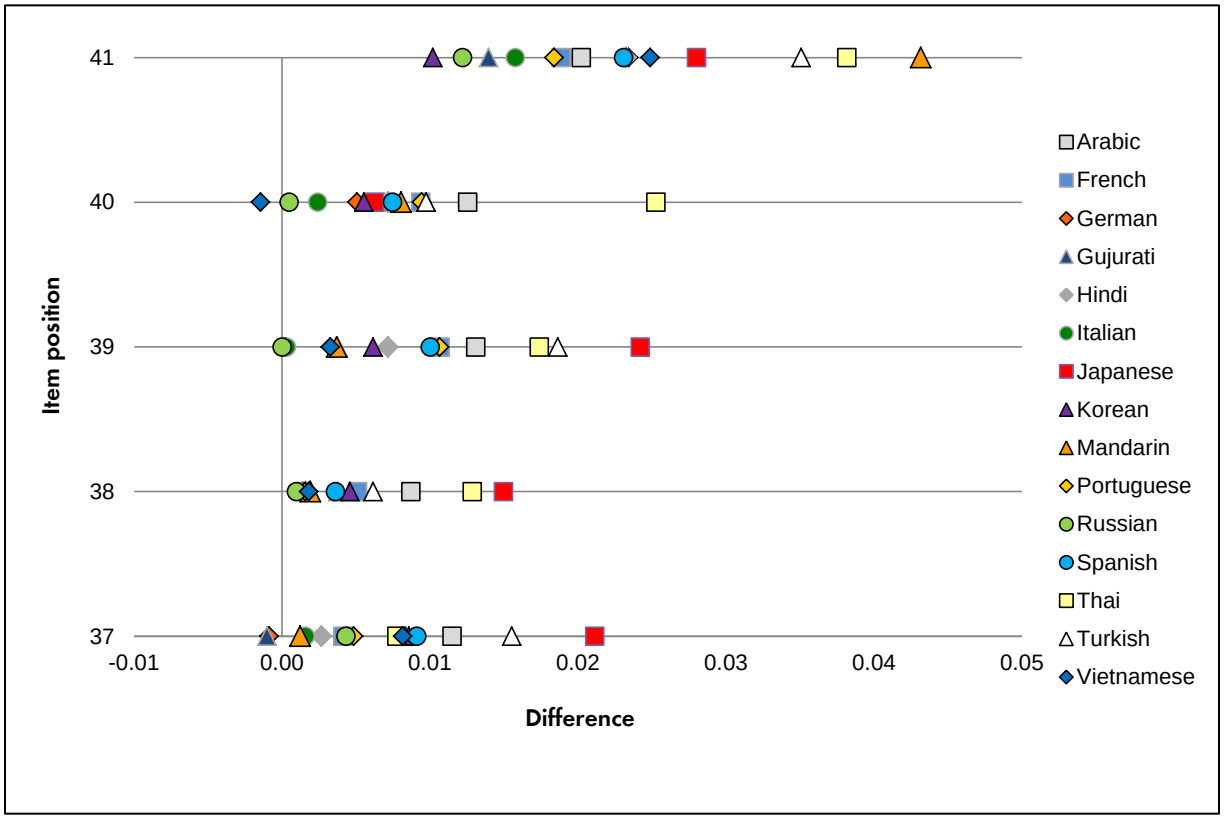
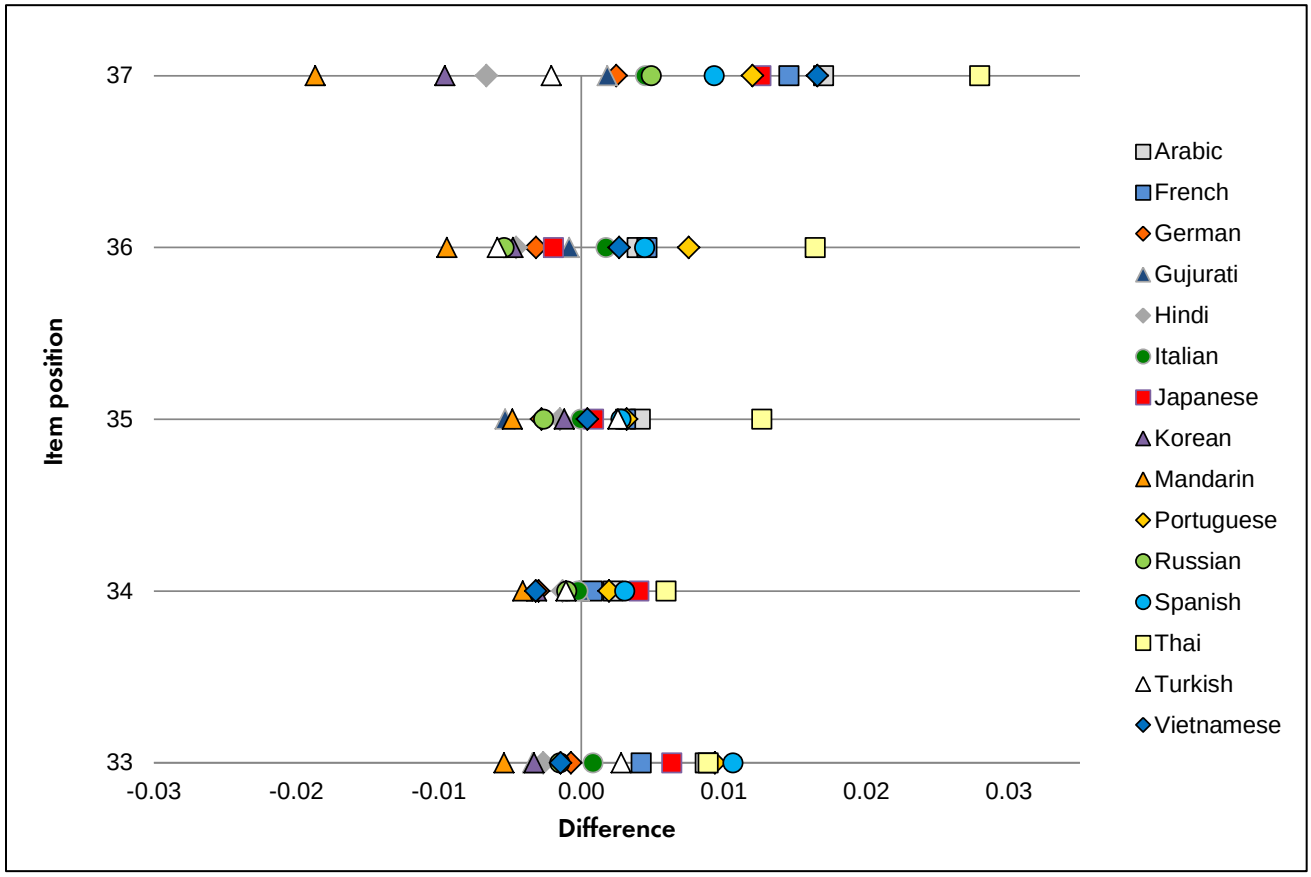


Figure 4. Difference in Proportion of Examinees Guessing on Quant by Item Position Compared to English



Differential speededness using the standardization approach takes into account ability differences across groups. Table 3 lists the average test scores across the language groups, and the differential not reached rates are illustrated in Figures 5 and 6. Differential rates of not reaching the last 10 items were examined relative to final scaled scores. According to Dorans, Schmitt, and Bleistein (1988) absolute differences less than 0.05 are negligible and those above 0.1 could cause concern. None of the comparisons yielded differences greater than 0.1. Only native Korean speakers observed differences in proportions not reaching the last 10 or more items greater than 0.05, with a difference in proportion of 0.09 not reaching item 32 or higher (out of 41) on the verbal section versus native English speakers, and a difference in proportion 0.06 not reaching item 28 or higher (out of 37) on the quantitative section. Because these two values are at the limit of the items studied (the 10th item), they combine all those dropping out up to that point, which from the previous analyses seem to be unusually high for the Korean group. With the exception of those two values, the standardization approach, conditioning on ability, suggests there is no differential speededness across groups. Contrasting the standardization results with the other results for speededness and time pressure indicates that, when conditioned on ability, even the differences for the apparent outlier Korean language group compared to the English language group are negligible.

Table 3. Mean (SD) Scaled Scores by Language				
Language	**N**	**Verbal**	**Quant**	**Total**
English	114,042	30.80 (8.00)	32.88 (10.09)	537.87 (117.94)
Arabic	4,299	20.29 (8.71)	30.54 (11.26)	440.12 (130.44)
Mandarin	26,881	24.22 (8.93)	45.84 (6.58)	584.24 (104.10)
German	3,312	30.16 (8.54)	37.87 (8.19)	567.61 (109.81)
Spanish	7,557	25.96 (8.38)	32.60 (11.07)	499.68 (124.84)
French	3,925	27.61 (8.98)	36.61 (10.20)	540.23 (126.87)
Guajarati	1,853	21.61 (9.69)	33.57 (12.08)	471.91 (147.31)
Hindi	12,464	27.19 (8.45)	42.41 (8.64)	580.01 (114.78)
Italian	1,100	29.49 (8.25)	38.62 (8.73)	568.40 (109.14)
Japanese	1,504	21.11 (7.97)	41.93 (8.21)	527.33 (103.65)
Korean	4,036	23.41 (8.34)	43.72 (7.91)	561.18 (105.76)
Portuguese	1,868	27.49 (9.09)	36.49 (10.46)	538.53 (129.84)
Russian	3,400	27.55 (9.23)	37.70 (9.48)	547.16 (122.24)
Thai	1,514	19.40 (7.73)	38.63 (9.55)	489.14 (111.48)
Turkish	1,623	22.32 (8.60)	42.59 (8.20)	541.87 (111.21)
Vietnamese	1,466	22.88 (8.51)	37.05 (10.25)	506.79 (118.34)

Figure 5. Differential Not Reached Rates Compared to English by Items Positions for Verbal

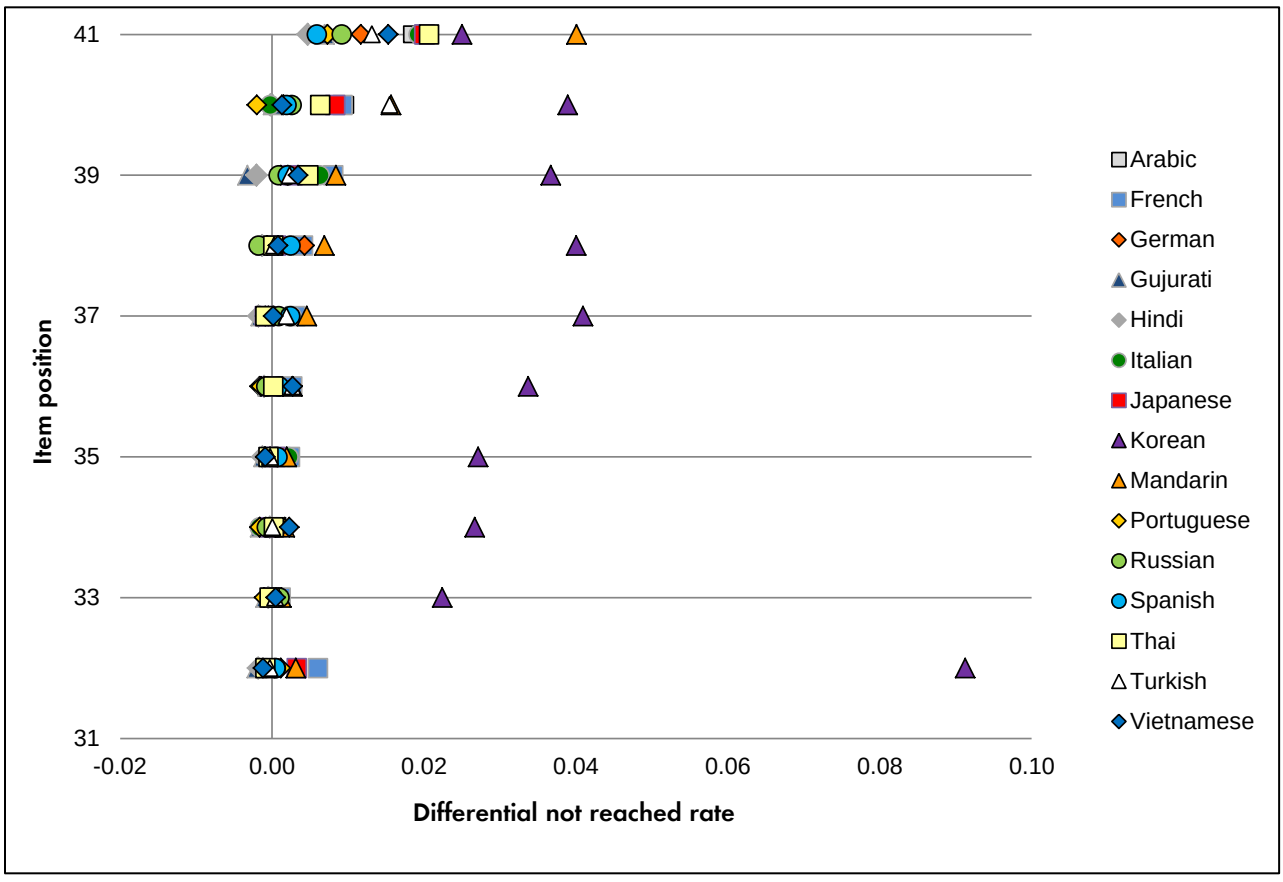
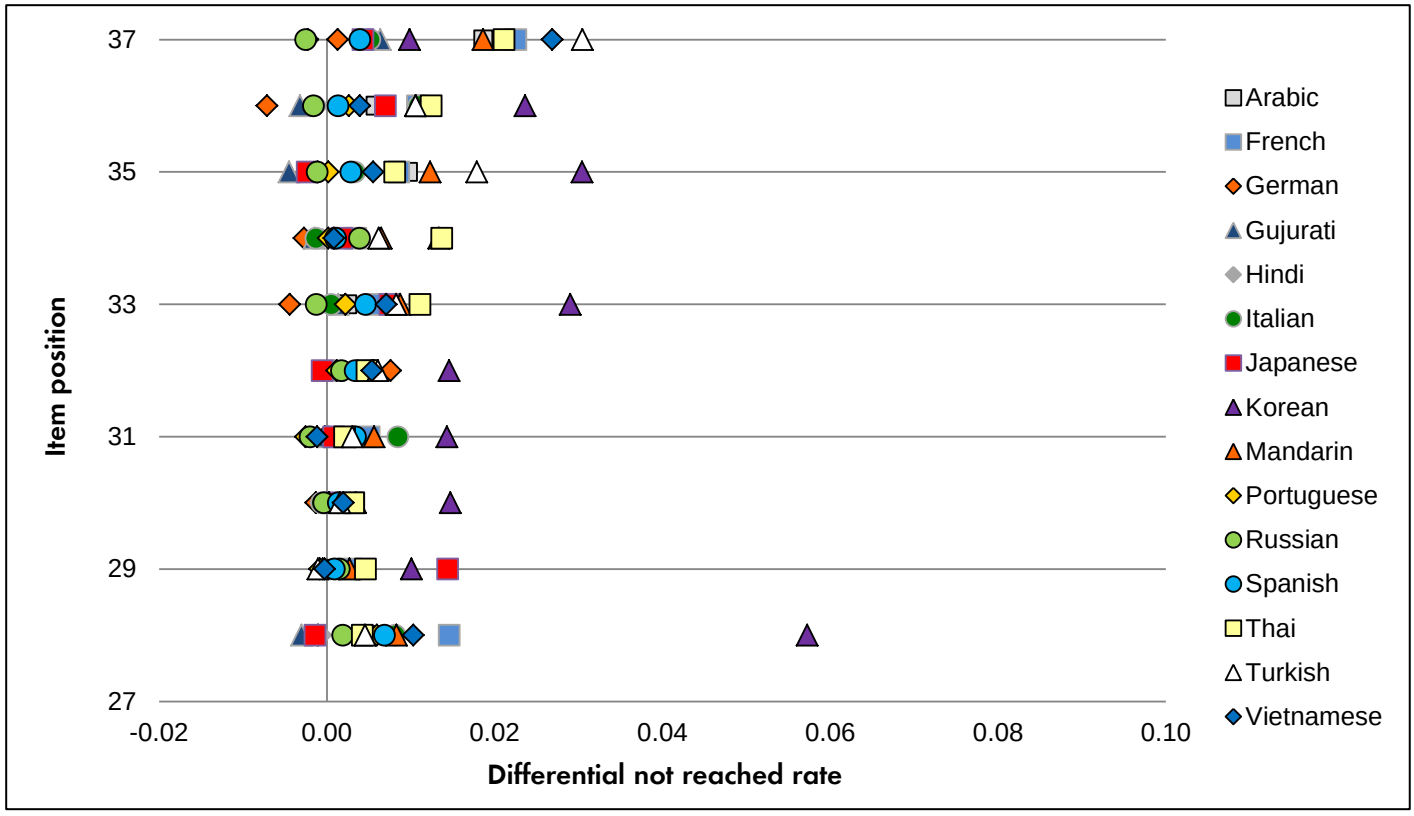


Figure 6. Differential Not Reached Rates Compared to English by Item Position for Quant



Discussion

The study revealed differences in speed components based on the native languages of examinees taking the GMAT exam. In particular, those whose native language was listed as Korean had far greater time difficulties than any of the other groups. There are several possible explanations for differences in time across groups. Differences in the grammatical structure of some of the Asian languages (especially the Ural-Altaic language family), such as Korean and Japanese, may contribute to differences observed in time pressure. One issue with the current study is that there is no measure of English proficiency for the examinees. The question would be whether the time pressure is a result of lack of English proficiency, which could be considered construct relevant. Previous research suggests that different language or cultural groups have different test-taking strategies. Defibaugh (2010) reported that certain non-native English language groups, such as Korean and Japanese, were statistically significantly more likely to repeat taking the GMAT exam. It may be that some groups take the actual test as a practice strategy and may not have exhibited full effort. Research on guessing behaviors in different cultures may provide information about why some candidates take longer or omit items at the end (Emengou & Childs, 2005).

Future research on differential speededness by language groups should include information such as scores from an English language test in order to disentangle English proficiency from speed concerns. It should be noted that the ability estimates used in the standardization approach were the scaled scores, which included the possibly confounding effects of the time limit. Ideally, further research in this area should include an independent measure of ability. Additional validity research including language variables can serve to alleviate concerns about whether time limits disadvantage certain groups. The current study illustrates that distinct differences exist among language groups, suggesting that comparing all non-English-speaking groups to native English speakers

may mask possible differences. These differences and others, such as repeat test taking, demonstrate that there are many challenges that should be considered when evaluating tests to be used for global purposes.

Contact Information

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

Notes

An earlier version of this paper was presented at the annual meeting of the National Council of Measurement in Education, April 7–11, 2011, New Orleans, LA.

The views and opinions expressed in this article are those of the authors and do not necessarily reflect those of the Graduate Management Admission Council®.

References

Bridgeman, B. (2004). *Speededness as a threat to construct validity*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Bridgeman, B. & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement, 41*, 137–148.

Chang, S. (2007). *Computerized adaptive test item response times for correct and incorrect pretest and operational items: Testing fairness and test-taking strategies* (Doctoral dissertation). University of Nebraska-Lincoln, Lincoln, NE.

Defibaugh, C. (2010). *The who and why: Repeat testing patterns around the world on the GMAT exam*. Presentation at the meeting of the International Test Commission, July 19–21, 2010, Shatin, Hong Kong.

Dorans, N., Schmitt, A., & Bleistein, C. (1988). *The standardization approach to assessing differential speededness*. Research Report 88–31, Princeton, NJ: Educational Testing Service.

Emengou, B. & Childs, R. (2005). Curriculum, translation, and differential item functioning of measurement and geometry items. *Canadian Journal of Education, 28*, 128–146.

Evans, F. & Reilly, R. (1973). A study of speededness as a source of test bias. *Journal of Educational Measurement, 9*, 123–131.

Guo, F. (2011). *Expected latency and automated test assembly*. Presentation at the annual meeting of the Association of Test Publishers, February 28–March 2, 2011, Phoenix, AZ.

Lawrence, I. (1993). *The effect of test speededness on subgroup performance*. Research Report 93–49, Princeton, NJ: Educational Testing Service.

Lord, F. & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lu, Y. & Sireci, S. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice, 26(4)*, 29–37.

Mislevy, R. & Wu, P. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing*. Research Report 96–30-ONR, Princeton, NJ: Educational Testing Service.

Pennock-Román, M. (1992). Interpreting test performance in selective admissions for Hispanic students. In Geisinger, K. (Ed) *Psychological testing of Hispanics*, APA science volumes (pp. 99–135). Washington, DC: American Psychological Association.

Peterson, N. (1993). *Review of issues associated with speededness of GATB tests*. Washington, DC: American Institutes for Research.

Rindler, S. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement, 16*, 261–270.

Schmitt, T., Sass, D., Sullivan, J., & Walker, C. (2010). A Monte Carlo simulation investigating the validity and reliability of ability estimation in item response theory with speeded computer adaptive tests. *International Journal of Testing, 10*, 230–261.

Sireci, S. & Talento-Miller, E. (2006). Evaluating the predictive validity of Graduate Management Admission Test scores. *Educational and Psychological Measurement, 66*, 305–317.

Talento-Miller, E. (2008). Generalizability of GMAT® validity to programs outside the U.S. *International Journal of Testing, 8*, 127–142.

Talento-Miller, E. & Guo, F. (2009). *Guess what? Score differences with rapid replies versus omissions on a computerized adaptive test*. GMAC Research Report Series RR-09-04. McLean, VA: Graduate Management Admission Council.

