ETS GRE®

Assess Ability. Predict Performance.

ETS GRE® Board Research Report
ETS GRE®–14-02
ETS Research Report No. RR–14-25

Calculator Use on the *GRE®* revised General Test Quantitative Reasoning Measure

Yigal Attali

December 2014

The report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

****

****

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

GRE-ETS

PO Box 6000

Princeton, NJ 08541-6000

USA

To obtain more information about GRE programs and services, use one of the following:
Phone: 1-866-473-4373
(U.S., U.S. Territories*, and Canada)
1-609-771-7670
(all other locations)
Web site: www.gre.org

*America Samoa, Guam, Puerto Rico, and US Virgin Islands

GRE Board Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Calculator Use on the *GRE*® revised General Test Quantitative Reasoning Measure

Yigal Attali

Educational Testing Service, Princeton, NJ

Previous research on calculator use in standardized assessments of quantitative ability focused on the effect of calculator availability on item difficulty and on whether test developers can predict these effects. With the introduction of an on-screen calculator on the Quantitative Reasoning measure of the *GRE*® revised General Test, it is possible to explore calculator use in more detail. This study investigates calculator usage for the GRE examinee population across examinee and item characteristics. Results suggest that calculator usage in the GRE is very common (75% of responses used a calculator) but is less prevalent for low and high ability examinees and for male, Black, and Hispanic examinees. Items associated with lower levels of calculator usage tended to be more difficult, required less time to answer, and did not present a real-world application. Quantitative comparison items were also associated with less calculator usage, but other item type and content distinctions showed small differences in usage. Most items exhibited a positive relation between calculator usage and response accuracy (after controlling for examinee ability), that is, responses that used the calculator were more likely to be correct. Items that did not show an association between calculator use and accuracy tended to be easier and had less overall usage of the calculator.

Keywords Calculators; mathematics; standardized testing; GRE

doi:10.1002/ets2.12025

Student use of calculators in mathematics instruction and assessment is now ubiquitous. Calculators are now permitted in nearly all mathematics courses and are even required in many high school courses. Calculator use is allowed in all major standardized tests, including the ACT, *SAT*®, and *GRE*® assessments. The National Assessment of Educational Progress (NAEP) allows the use of a calculator in approximately one third of the assessment at each grade level. A graphing calculator is considered an integral part of the *AP*® calculus course and is allowed on the AP calculus exams.

However, calculator use has been a controversial issue for mathematics educators since calculators became widely available to students in the late 1970s. It is not uncommon to find educators who encourage the complete abolishment of paper-and-pencil computations (e.g., Ralston, 1999, p. 2), while others recommend extremely limited use for calculators (e.g., Mackey, 1999, p. 3).

In the United States, the National Council of Teachers of Mathematics (NCTM) was an early advocate for calculator use in the classroom with their recommendation that "mathematics programs take full advantage of the power of calculators and computers at all grade levels" (NCTM, 1980, p. 8) to enhance students' understanding and use of numbers and operations. This position was reaffirmed several times since (e.g., NCTM, 1989), although more recent position statements have also emphasized the importance of computation skills, including estimation and mental mathematics, and the appropriate use of the calculator (NCTM, 2005).

Research has largely supported this position. Hembree and Dessart (1986, 1992) conducted a meta-analysis of the effects of calculator use in mathematics programs. In general, they found that the availability of calculators during instruction had no significant effect on student performance but had a positive influence on students' attitudes toward mathematics (compared to students with no availability to calculators). A more recent meta-analysis by Ellington (2003) separated results into studies that allowed calculators during testing versus those that did not. When calculators were included in instruction but not testing, operational and problem-solving skills improved with no changes in computational and conceptual skills. When calculators were part of both testing and instruction, all four of these skills improved (with small to moderate effect sizes). Students who used calculators while learning mathematics also reported more positive attitudes toward mathematics.

Corresponding author: Y. Attali, E-mail: yattali@ets.org

A number of studies also compared performance of students on quantitative tests with and without the availability of calculators (Loyd, 1991; Morgan & Stevens, 1991). A large-scale (7,000 examinees from 275 high schools) study conducted when the calculator was introduced for the SAT indicated a generally positive effect (items became easier) for students who were allowed to use calculators (Bridgeman, Harvey, & Braswell, 1995). Interestingly, a minority of items became harder with calculator availability. An example of such an item asked for the remainder when 63,383 is divided by 7. It is possible that the availability of the real number quotient through the calculator confused some examinees.

Another large-scale study was conducted in preparation for the introduction of calculators in the GRE revised General Test (Bridgeman, Cline, & Levin, 2008). In this study, 168 items were administered in six research forms either with or without an available calculator. On average, calculator availability increased percent correct by 1%. In addition, 9% of items showed an increase of at least 5% correct. The total effect on GRE scores was 12 points (*SD* was around 130), and there were no interactions with gender or ethnicity. Calculator use varied from 0% to 61% across items, with a median of 19%. For the 20 items with the largest calculator use, percent correct for those who chose to use a calculator was generally higher than for those who chose not to use it (by an average of 20%). However, note that this result is for a small subset of items and ability was not controlled. Finally, some items with calculator availability that were predicted (by judgments of test developers) to show a calculator effect, also showed lower average response times.

The previous studies examined the effects of calculator availability in experimental settings. However, once a high-stakes testing program makes calculators available to examinees, a different set of questions arise. These questions stem from the fact that although the calculator is available, examinees may or may not choose to use it for any given item on the test. Only one study evaluated actual calculator use on a high-stakes testing program. During a 1996 and 1997 administration of the SAT (2 and 3 years after a calculator was allowed on the test), examinees were asked whether and what kind (four-function, scientific, graphing, other) of calculator they brought to the test and in how many questions did they use the calculator (none, a few, about a third, about half, most). Scheuneman, Camara, Cascallar, Wendler, & Lawrence 2002, see also Wendler, Zeller, & Allspach, 2003) analyzed this data (from more than 200,000 examinees). Results showed that almost all examinees brought a calculator, but only 15% used it for most of the items. Girls used the calculator more often than boys, and White and Asian examinees used the calculator more often than other groups. Students who used the calculator on one third to one half of the questions performed better than those who used it more or less often. Calculator use during the test was not found to predict scores beyond background academic variables — only availability of a graphing calculator in the test and calculator access at home increased prediction accuracy. A differential item functioning (DIF) analysis with the responses to the three calculator questions as a basis for contrasting groups was also performed (e.g., contrasting examinees that answered calculator was used in no items versus all other examinees). Overall, 5 items in one sample and nine items in another sample (out of 60) were identified as showing DIF for the two samples, although one of five and four of nine favored no items for which the calculator was used.

Because the SAT is a paper-and-pencil test, measurement of calculator use in Scheuneman et al. (2002) was not precise. The purpose of this study is to explore more closely actual calculator usage in a high-stakes quantitative assessment that allows precise measurement of calculator usage for each item. The GRE revised General Test, launched in August 2011, allows the use of an on-screen calculator for the Quantitative Reasoning measure. The quantitative sections of the GRE test emphasize reasoning skills over computational skills. Nevertheless, many items require nontrivial complex computations to answer. The on-screen calculator is a basic four-function calculator with parentheses, square-root, and memory buttons. It also allows test takers to directly transfer the computed result to answer open-ended items.

Now that an on-screen calculator is available, examinees have to decide whether or not to use the calculator for each item. The purpose of this study is to explore calculator usage on the test across examinees and items. What characteristics of examinees (ability, gender, and ethnicity) are related to calculator use? What characteristics of items (difficulty, typical response time, item type, and content classifications) are related to calculator use? Is response accuracy related to calculator use across items and after controlling for examinee ability?

Design

The GRE is a multistage test (MST). Examinees are administered two Verbal Reasoning and Quantitative Reasoning sections as well as one Analytical Writing section. The GRE is adaptive at the stage, not question, level, and the determination of the next set of questions an examinee receives is based on performance on an entire preceding stage. In addition, examinees are administered one variable section (either verbal or quantitative) that serves calibration or pretest purposes.

Quantitative sections are composed of 20 items and the content specification of the variable section is similar to the operational sections. The quantitative sections are composed of several item types. In addition to regular (single-selection) multiple-choice items, the test includes multiple-selection multiple-choice items (where more than one option can be correct), quantitative comparison (QC) items (in which examinees need to determine whether one of two quantities is greater than the other, they are equal, or their relationship cannot be determined), and numeric entry items (open-ended items with a number as correct answer). Variable sections are administered to representative samples of several thousand examinees. For this study, a random sample of 1,000 domestic examinees from each of 20 quantitative calibration sections was selected for analysis. The sections were administered in the fall of 2012, more than a year after the introduction of the GRE, so that examinees had time to get used to the available calculator. Although all examinees in each of the 20 samples were administered the same variable section, they were administered different operational sections.

Information available for each examinee included the final scaled quantitative score, gender, and ethnicity. Information available for each response included response accuracy, response time, and whether or not the calculator was used for the item. Content classifications for each item include item type (single-selection multiple-choice, multiple-selection multiple-choice, QC, and numeric entry), content area (algebra, arithmetic, data analysis, and geometry), and a subclassification of items to real (word problems with applied contexts) or pure (non-word problems).

Results

Examinee-Related Analyses

Figure 1 presents the distribution of overall calculator use for the operational sections, measured by summing the number of items where a calculator was used (with a possible range of 0–40). The figure shows that the calculator is used often by most examinees ($Md = 32$, $M = 30$, $SD = 10$). A similar distribution is found for the variable section. The internal consistency of calculator use is very high, with Cronbach's alpha coefficients of .925, .915, and .916 for the first and second operational sections and for the variable section, respectively, and .950 for the two combined operational sections.

The correlation between operational calculator use and quantitative (Q) scaled score[1] is .11, but as Figure 2 shows, the relation between these two measures is not linear (for cubic, or third-degree polynomial regression, $R = .28$). Examinees with the lowest 10–20% of Q scores show lower calculator use, as well as examinees with the highest 5% of Q scores.

To examine possible differences in calculator use across gender and race/ethnicity (White, Black, Asian, and Hispanic), a two-way analysis of variance was performed with operational calculator use as the dependent measure. The gender effect was significant: $F(1, 17{,}249) = 62.0$, $p < .01$, $\eta_p^2 = .0036$, with higher calculator use for women ($M = 30.3$, $SD = 9.4$) than for men ($M = 28.4$, $SD = 10.4$). The race/ethnicity effect was also significant, $F(3, 17{,}249) = 41.3$, $p < .01$, $\eta_p^2 = .0071$, with higher calculator use for White ($M = 29.9$, $SD = 9.6$) and Asian ($M = 30.1$, $SD = 10.1$) examinees than for Black ($M = 26.5$, $SD = 11.0$) and Hispanic ($M = 28.9$, $SD = 10.0$) examinees, as revealed by post hoc comparisons. The interaction effect was not significant.

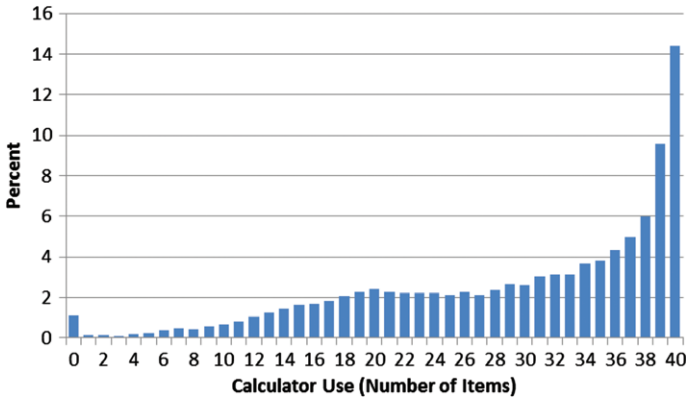

Figure 1 Histogram of calculator use in operational sections.

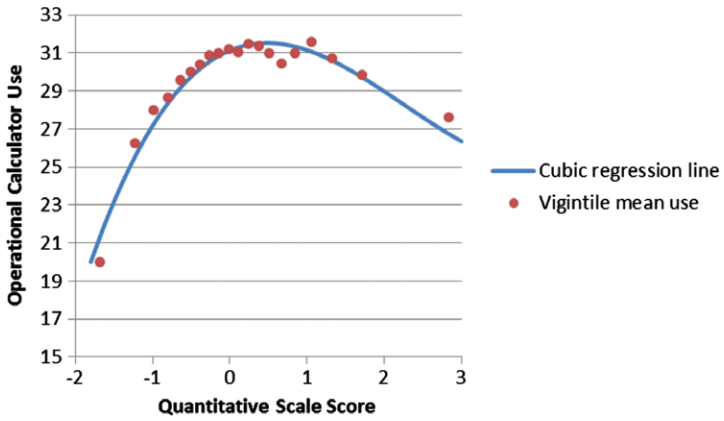

Figure 2 Relation of quantitative scores with operational calculator use.

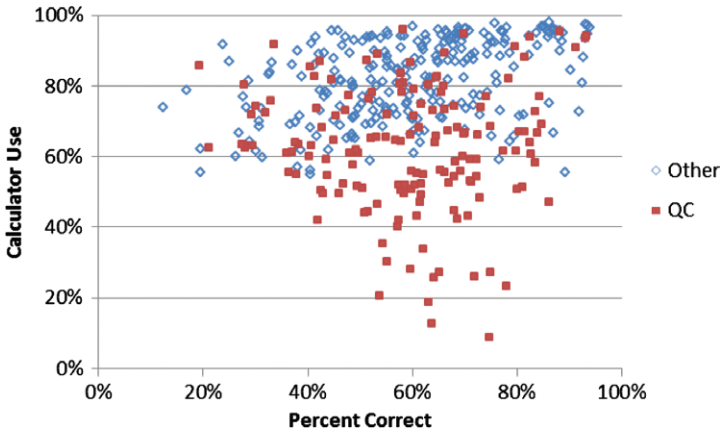

Figure 3 Item easiness and calculator use in qualitative comparisons (QC) and non-QC (other) items.

Item-Related Analyses

To examine possible differences in overall calculator use across different item characteristics, the percentage of examinees that used a calculator for each of the 400 items in the variable sections was computed. First, the relation between calculator use and item difficulty (percent correct) was analyzed. An initial inspection of the data showed a different pattern of results for QC and other items, which is presented in Figure 3. As shown in the figure, QC items are associated with lower levels of calculator use and a lack of any relationship between calculator use and item easiness ($r = -.01$), whereas other items are associated with higher levels of calculator use and a moderate relation between calculator use and item easiness ($r = .42$).

Figure 4 shows a moderate relationship between average response time and calculator use ($r = .36$). An inspection through Figures 3 and 4 of the items with the lowest levels of calculator use shows that they are all QC items and have short average response time. The QC item type was designed to be answered quickly and with minimal computational burden, and these results show that test developers are able to achieve this goal with some of these items.

In contrast to item easiness and average response time, item discrimination (measured by the Biserial statistic) showed no relationship with calculator use, both for QC items ($r = .04$, $p = .60$) and non-QC items ($r = -.10$, $p = .10$).

To examine possible differences in calculator use across the content classifications (item type, content area, and real versus pure problems), an analysis of covariance was performed with item calculator use as the dependent measure and item easiness as covariate. This analysis was performed only for non-QC items because the results in the preceding section have shown that item easiness is not linearly related to calculator usage for QC items.

The item type effect (after controlling for item easiness) was barely significant, $F(2, 238) = 3.9$, $p = .02$, $\eta_p^2 = .032$. Post hoc comparisons using the Tukey HSD test found only one significant difference between multiple-selection items (adj. $M = .76$) and numeric entry items (adj. $M = .83$), with single-selection items between the two other item types

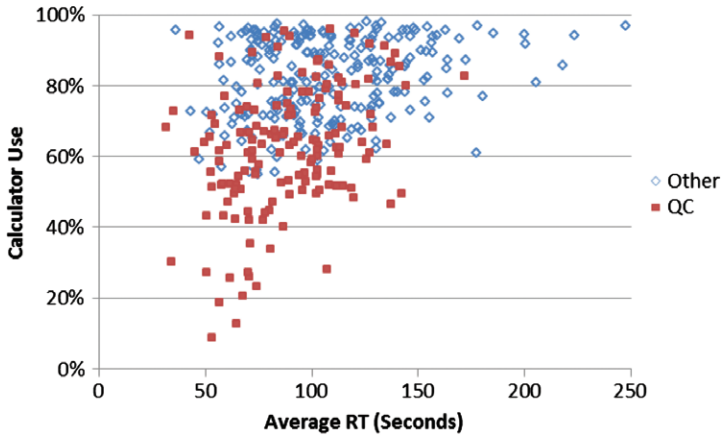

Figure 4 Item average response time (RT) and calculator use in qualitative comparisons (QC) and non-QC (other) items.

(adj. $M = .82$). The content area effect was also barely significant, $F(3, 238) = 3.2$, $p = .02$, $\eta_p^2 = .039$, with adjusted means of .77 for algebra, .79 for geometry, .81 for data, and .83 for arithmetic. Post hoc comparisons indicated that only the difference between the algebra and arithmetic adjusted means were significant. This finding is in accordance with expectations, since arithmetic items are expected to require more computations. The real versus pure effect was also significant: $F(3, 238) = 46.3$, $p < .01$, $\eta_p^2 = .163$, with lower calculator use for pure items (adj. $M = .75$) than for real items (adj. $M = .85$). This finding is in accordance with expectations as well, since real items often involve real life problems.

Relation of Calculator Use and Response Accuracy

Next, we examined the relationship between calculator use and correctness of response. For each item, it was possible to examine the direction and strength of association between these two binary variables across all examinees. However, to avoid possible confounding with respect to examinee ability, a Mantel-Haenszel (MH) analysis was used. For each item, examinees were stratified into 10 equal-sized groups based on their quantitative score. An MH analysis then evaluates the strength of association by estimating the common odds ratio, across strata, for using the calculator versus not using it. When the odds ratio is higher than 1, the odds (and probability) of using the calculator are higher for correct responses than for incorrect responses, indicating a positive association between calculator use and correctness of response. When conducting an MH analysis, the Breslow-Day statistic tests whether the odds ratios across strata (ability levels in this case) are homogeneous.

In the sample of 400 items, a range of common odds ratios for calculator use and correctness of response was observed ($Mdn = 1.36$, $M = 2.12$, $SD = 2.37$).[2] A Wilcoxon signed-ranks test indicated that the median common odds ratio was greater than 1, $Z = 12.81$, $p < .01$, with 76% of items showing a common odds ratio greater than 1. To illustrate the size of the median effect, consider that for an item with average difficulty (60% correct) and average calculator use (75% use), the median odds ratio of 1.36 would correspond to 62% correct for examinees who used the calculator versus 54% correct for examinees who did not use the calculator (controlling for examinee ability). Alternatively, it would correspond to 71% calculator use for wrong answers and 77% calculator use for correct answers.

Figures 5 and 6 show a positive relationship between strength of association (between calculator use and correctness) and both item easiness ($r = .41$) and item calculator use ($r = .56$), indicating that easier items and items with higher levels of calculator use also tend to exhibit stronger association between calculator use and correctness. These relations seem to hold for both QC and non-QC items. In these analyses, the natural logarithm of the common odds ratio was used[3] so that positive values indicated positive associations and vice versa.

To examine possible differences in association of calculator use with item correctness across the content classifications (item type, content area, and real versus pure problems), an analysis of covariance was performed on all items with logarithm of common odds ratio as the dependent measure and item easiness and item calculator use as covariates. None of the three main effects was statistically significant (after controlling for item easiness).

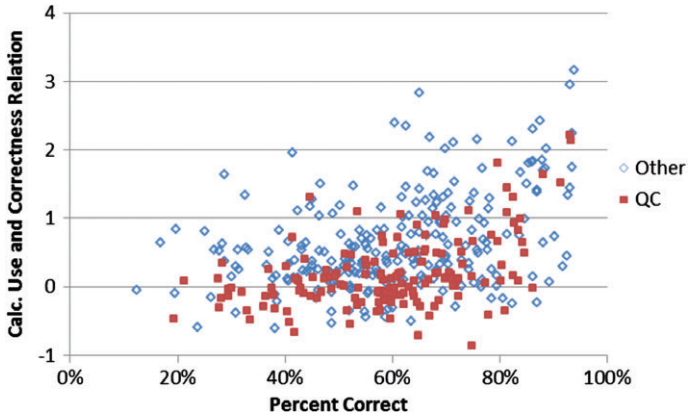

Figure 5 Item easiness and association of calculator use with correctness.

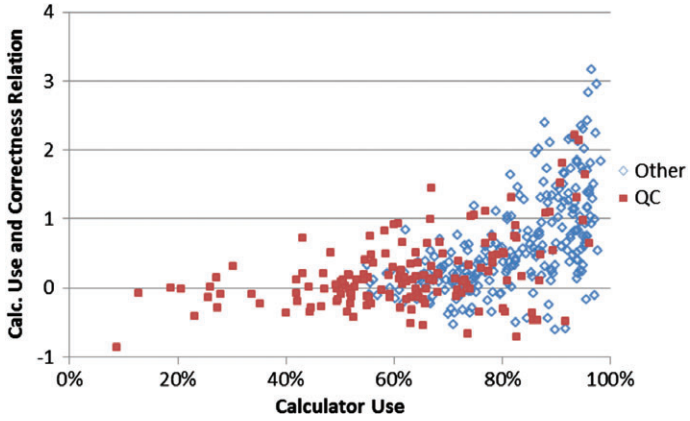

Figure 6 Item calculator use and association of calculator use with correctness.

Conclusions

This study sought to explore calculator use on a high-stakes quantitative test. The quantitative section of the GRE test emphasizes reasoning skills over computational skills. Nevertheless, many items require nontrivial computations to answer. Therefore, in principle, the availability of an on-screen calculator should enhance the construct validity of the assessment as the computational burden on students is eased, allowing them to concentrate on the more central problem-solving skills. A calculator can help avoid trivial computational errors and can speed the process of solving a problem. However, the use of a calculator requires skill and judgment, and the decision to use a calculator may affect the approach taken for solving the problem.

Results of this study indicate that calculator usage in the GRE is very common, as the calculator was used in around 75% of item responses. This finding suggests that examinees use the calculator to verify even simple calculations. However, it may also be the case that the typical computational burden in answering items is not negligible (for very few items, calculator usage is below 50%). Test developers may feel less constrained in writing items that need some calculations to answer, now that a calculator is available for examinees.

Examinee-related analyses reveal that measurement of calculator usage (through the number of items for which the calculator was used) is highly reliable (internal consistency measure of .95). Interestingly, the relationship between usage and quantitative scores is nonlinear, with both lower ability examinees and higher ability examinees (the highest decile) showing lower calculator usage. Lower ability examinees may use the calculator less often because they are unsure of the calculation to perform. Higher ability examinees may use the calculator less often because they may feel more confident in performing some calculations mentally. Similarly to what Scheuneman et al. (2002) found for the SAT, women used the calculator more often than men ($d = .20$), and White examinees used it more often than both Black

examinees ($d = .34$) and Hispanic examinees ($d = .10$), although these results may have been confounded by examinee ability.

Item-related analyses show that, for non-QC items, easier items tend to show higher calculator use. This finding may result from a tendency of harder items to be less computational, but it may also be the case that examinees who do not know how to solve a problem often would not know what the calculations are that should be performed (similarly to what was suggested for low ability students). Results also show that items requiring longer time to answer also tend to show higher calculator use, as expected from the intuition that computational problems require more time than conceptual problems. Similar reasoning can explain the differences between item types and content specifications. Lower calculator use was found for QC items that were designed to require less computations and time to answer. Similarly, arithmetic items showed higher calculator use than algebra items (83% vs. 77% use), and real items showed higher calculator use than pure items (85% vs. 75% use). Both arithmetic and real items are expected to require more computations than their counterparts. However, the differences that were found are not large, underscoring the general tendency of widespread calculator use.

Analyses of the relation between calculator use and response accuracy show a general positive association between them, indicating that, for most items, examinees who use the calculator are more likely to answer correctly than examinees (with the same quantitative score) who do not use the calculator. This association is generally stronger for easier items and for items with higher calculator use. We cannot infer from these correlational results about the causal relationship between calculator use and success in solving the problems. One causal mechanism that was invoked to provide a possible explanation for some of the previous findings in this paper is that understanding affects calculator use. That is, examinees who do not understand the question and the computations required to reach an answer will tend not to use the calculator. A different mechanism would postulate that calculator use affects correctness. That is, examinees who do not use the calculator might make more computational mistakes that lead to incorrect responses. Of course, these two mechanisms (and possibly others) could operate at the same time for different items and different examinees.

However, in both types of explanations, calculator-supported computations are needed to supplement student understanding. In this respect, as well as the widespread use of the calculator, the results of this study support the availability of the on-screen calculator in the GRE.

Notes

1 The scaled score is the continuous equated GRE score with an approximate mean of 0 and standard deviation of 1. These scores are then transformed to the reported scores on a whole-number scale of 130 – 170.
2 The Breslow-Day statistic was significant (at the 1% level) only in 3% of the items, indicating homogeneity of association across ability levels for almost all items.
3 This was done to achieve the same scale for negative and positive associations.

References

Bridgeman, B., Cline, F., & Levin, J. (2008). *Effects of calculator availability on GRE quantitative questions* (Research Report No. RR-08-31). Princeton, NJ: Educational Testing Service.

Bridgeman, B., Harvey, A., & Braswell, J. (1995). Effects of calculator use on scores on a test of mathematical reasoning. *Journal of Educational Measurement*, *32*, 323 – 340.

Ellington, A. J. (2003). A meta-analysis of the effects of calculators on students' achievement and attitude levels in precollege mathematics classes. *Journal for Research in Mathematics Education*, *34*, 433 – 463.

Hembree, R., & Dessart, D. J. (1986). Effects of hand-held calculators in precollege mathematics education: A meta-analysis. *Journal of Research in Mathematics Education*, *17*, 83 – 99.

Hembree, R., & Dessart, D. J. (1992). Research on calculators in mathematics education. In J. T. Fey (Ed.), *Calculators in mathematics education* (pp. 23 – 32). Reston, VA: National Council of Teachers of Mathematics.

Loyd, B. H. (1991). Mathematics test performance: The effects of item type and calculator use. *Applied Measurement in Education*, *4*, 11 – 22.

Mackey, K. (1999). Do we need calculators? In Z. Usiskin (Ed.), *Mathematics education dialogues* (p. 3). Reston, VA: National Council of Teachers of Mathematics.

Morgan, R., & Stevens, J. (1991). *Experimental study of the effects of calculator use on the Advanced Placement Calculus Examinations* (Research Report No. RR-91-05). Princeton, NJ: Educational Testing Service.

National Council of Teachers of Mathematics. (1980). *An agenda for action: Recommendations for school mathematics in the 1980s*. Reston, VA: Author.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2005). *Computation, calculators, and common sense*. Reston, VA: Author.

Ralston, A. (1999). Let's abolish pencil-and-paper arithmetic. In Z. Usiskin (Ed.), *Mathematics education dialogues* (p. 2). Reston, VA: National Council of Teachers of Mathematics.

Scheuneman, J. D., Camara, W. J., Cascallar, A. S., Wendler, C., & Lawrence, I. (2002). Calculator access, use, and type in relation to performance on the SAT I: Reasoning test in mathematics. *Applied Measurement in Education*, *15*, 95–112.

Wendler, C., Zeller, K., & Allspach, J. (2003, April). *The impact of calculator preference and use on performance on a mathematics reasoning test.* Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Suggested citation:

Attali, Y. (2014). *Calculator use on the GRE® revised General Test quantitative reasoning measure* (GRE Board Research Report No. 14-02, ETS Research Report No. RR-14-25). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12025

Action Editor: Brent Bridgeman

This report was reviewed by the GRE Technical Advisory Committee and the Research Committee and Diversity, Equity and Inclusion Committee of the GRE Board.