DOCUMENT RESUME

ED 382 656 TM 023 090

AUTHOR	Lawrence, Ida M.; Dorans, Neil J.
TITLE	Optional Use of Calculators on a Mathematical Test: Effect on Item Difficulty and Score Equating.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-94-40
PUB DATE	Aug 94
NOTE	23p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Calculators; College Entrance Examinations; *Difficulty Level; *Equated Scores; Estimation (Mathematics); High Schools; *High School Students; *Mathematics Tests; *Scoring; Test Items
IDENTIFIERS	Anchor Tests; *Scholastic Aptitude Test

ABSTRACT

This paper describes findings from two studies involving optional use of calculators on Scholastic Aptitude Test (SAT) mathematical items. The first study looked at the effects of calculator use on estimates of item difficulty. The second study looked at the effect of calculator use on equating results from an anchor test design. Study 1, involving 46,637 students using calculators and 45,765 without calculators, looked at data on specific items that became inappropriate for a test that permits calculators because the skills measured by the item administered with a calculator are quite different when the item is administered without a calculator. Study 2, involving 1,900 high school juniors with calculators and 1,860 without, showed that, because the use of a calculator sometimes makes items easier, any equating design that utilizes an anchor test design needs to ensure that the anchor test is administered under the same conditions, i.e. with a calculator or without a calculator. A solution for making this adjustment with a special equating study is described. One figure and five tables present study findings. (Contains 9 references.) (SLD)

RR-94-40

RESEARCH REPORT

OPTIONAL USE OF CALCULATORS ON A MATHEMATICAL TEST: EFFECT ON ITEM DIFFICULTY AND SCORE EQUATING

Ida M. Lawrence
Neil J. Dorans

Educational Testing Service
Princeton, New Jersey
August 1994

Optional Use of Calculators on a Mathematical Test: Effect on Item Difficulty and Score Equating

Ida M. Lawrence [1,2]

Neil J. Dorans

[1]The authors thank Brent Bridgeman, Linda Cook, Howard Everson, Skip Livingston, Gretchen Rigol, and Mike Schaffer for their helpful comments on an earlier draft of this paper.

[2]Support for this research was provided by the College Board and Educational Testing Service.

Abstract

This paper describes findings from two studies involving optional use of calculators on SAT mathematical items. The first study looked at the effects of calculator use on estimates of item difficulty. The second study looked at the effect of calculator use on equating results from an anchor test design. Study 1 looked at data on specific items that become inappropriate for a test that permits calculators because the skills measured by the item administered with a calculator are quite different when the item is administered without a calculator. Study 2 showed that, because the use of a calculator sometimes makes items easier, any equating design that utilizes an anchor test design needs to ensure that the anchor test is administered under the same condition, i.e., with a calculator or without a calculator. A solution for making this adjustment with a special equating study is described.

Optional Use of Calculators on a Mathematical Test: Effect on Item Difficulty and Score Equating

The SAT has been in use since 1926, and calculators and other aids have never been permitted on the test. However, beginning with the October 1993 administration of the Preliminary Scholastic Assessment Test/National Merit Scholarship Qualifying Tests (PSAT/NMSQT) and the March 1994 administration of the Scholastic Assessment Test (SAT), students have the option to use calculators on the mathematical sections of the tests. This change in policy reflects the opinion of the National Council of Teachers of Mathematics (1989) that calculators be integrated into the teaching and testing of mathematics. Under the planned policy, students may bring to the test and use any four-function, scientific, or graphing calculator. They will not be permitted to use "hand-held" minicomputers, pocket organizers, or lap-top computers.[1] See Rigol (1993) for a description of the rationale for this policy, and also a discussion of its associated advantages and disadvantages.

Several studies have shown that calculator use effects test performance (e.g., Bridgeman, Harvey & Braswell, 1992; Cohen & Kim, 1992; Morgan & Stevens, 1991; Loyd, 1991). The purpose of this paper is to describe findings from two studies involving optional use of calculators on SAT mathematical items. The first study looked at the effects of calculator use on estimates of item difficulty. The second study looked at the effect of calculator use on equating results from an anchor test design. Each study was based on separate data collection situations involving different sets of items and examinee groups.

---

[1] Details concerning the policy on calculator use are specified in a "*Q and A for Calculator Policy*", The College Entrance Examination Board, 1992.

Study 1

The data described below serve as an example of a context effect due to optional use of calculators on SAT mathematical items. A large shift in item difficulty values could be suggestive of a shift in the construct being measured by the new test. Bridgeman et al (1993) provide data on specific items in the SAT that become inappropriate for a test that permits calculators because the skills measured by the item administered with a calculator are quite different when the item is administered without a calculator. For instance, without a calculator a particular item may require estimation skills but with a calculator the estimation skills are no longer needed.

Data Source

Twenty experimental forms, each composed of unique sets of pretest items, were administered to high school classes. Sample sizes taking the various pretests ranged between 4,000 and 5,000 examinees. In addition to the pretest items, students also took a set of 25 SAT Quantitative Comparison (QC) items that was common across the twenty experimental forms. For ten of the experimental forms, students were permitted to use a calculator on the QC items and for ten of the experimental forms students were not permitted to use a calculator on the QC items. All students were permitted to use a calculator on the pretest items. There were 46,637 examinees in the calculator condition and 45,765 examinees in the no calculator condition.

Approximately 900 schools participated in the study. The twenty different experimental forms were spiraled within classroom. In the odd-numbered test booklets, a capital C appeared across the tops of all pages with questions. A note appeared at the beginning of the booklet stating YOU MAY USE A

CALCULATOR ON ALL QUESTIONS IN THIS TEST. In the even-numbered test booklets, capital Cs did not appear across the tops of pages with QC questions. A note appeared at the beginning of the booklet stating YOU MAY NOT USE A CALCULATOR ON QUESTIONS 1-25. Before question 26, the following statement appeared: YOU MAY USE A CALCULATOR ON QUESTIONS 26-35. The supervisor's manual indicated that some students would be asked to not use a calculator on questions 1-25. Also, supervisors were told to pass out the booklets in order, regardless of whether or not students had brought calculators to the testing.

Results of Group Level Analyses

The pretests taken with a calculator (n = 46,637) are easier, on average, than the pretests taken without a calculator (n = 45,765). Collapsed over all of the experimental forms, the mean raw score on the QC items is significantly higher for the calculator condition ($\underline{M}$ = 12.13, $\underline{SD}$ = 6.23) than for the no calculator condition ($\underline{M}$ = 11.23, $\underline{SD}$ = 6.07), $t=27.04$, $p< .001$).

Results of Item Level Analyses

Estimates of item difficulty are presented in terms of percentages of correct responses to each item (p). These estimates are shown in Table 1. The estimates of item difficulty are based on 39,260 examinees who took the items with a calculator and 38,645 examinees who took the items without a calculator[2]. Most of the items are unaffected by calculator use. However, for three of the items (17, 19, 22), there is a sizable calculator effect. These items are shown in Figure 1. Note that the computational load for all of these items is fairly high, so

[2]Samples used for the group analyses were reduced to smaller samples for the purposes of carrying out item level analyses.

it is not surprising that the items are easier under a calculator permitted condition. Similar results were reported by Cohen & Kim (1992).

Study 2

The purpose of score equating is to make scores on each new form of a test comparable to scores on other forms of the test. Typically, new forms of SAT-M are equated to old forms of SAT-M via an anchor test design. In this type of design the same anchor test is administered in the variable section of the new form and the old form. The anchor test section is composed of items that do not contribute to a student's score. The anchor test serves as a linkage between a new form and an old form, for which scores have previously been placed on the 200-to-800 College Board scale. Scores on the anchor test are used to measure and adjust for differences in the ability levels of the groups taking the new test and old test. The anchor test may be composed of items included in the test to be equated (internal anchor) or it may be administered as a separate section (external anchor).

In order to carry out an appropriate equating study, the anchor test needs to be administered under identical conditions when it accompanies the new test and the old test. This is because the anchor test procedure for collecting equating data requires that the items in the anchor test behave similarly for groups taking the new form and the old form. If SAT-M items administered with a calculator are generally easier than items administered without a calculator, and examinees were prohibited from using calculators on the anchor test accompanying the old test but were allowed to use a calculator on the anchor test accompanying the new test, the assumption of common material on the anchor test would not be met. This issue can be illustrated with data that show results of a methodologically correct equating (controlling for calculator use on the anchor

test) and a methodologically incorrect equating (not controlling for calculator use on the anchor test).

Data Source

A special administration of SAT-M items given under different conditions of calculator use provides a preliminary answer to the question concerning the potential effect of calculator use on score equating and score conversions. The sample consisted of high school juniors who indicated they planned to attend college. The sample was quite representative of students in the college bound senior SAT examinee population. Details concerning the experimental sample and how it was recruited are provided in Bridgeman et al (1993).

The first section in the administration was a 35-item section from a previously administered edition of the SAT; all students were prohibited from using a calculator on this section. An additional 70 items were also administered under calculator and no-calculator conditions. A subset of 36 items was selected from the full set of 70 items administered. These 36 items were intended to represent the kinds of items that would appear on an actual SAT. Scores on this 36-item test were placed on a 200-to-800 scale for illustrative purposes only[3].

The data from this experimental administration were used to carry out two equating analyses. The equatings were based on observed-score linear (Tucker) and curvilinear (chained equipercentile) anchor test models. Details concerning these equating methods can be found in Angoff (1984). Equatings were based on 1,900 examinees in the group allowed to use a calculator and 1,860 examinees in the group not allowed to use a calculator.

[3]This scaling would never be done in practice because the scale has too many possible scores (61) for a 36-item test and because the construct measured by the 36-item test may not be the same as the construct measured by the full-length test.

Analysis 1

A 36-item test given with a calculator was equated to the same 36-item test given without a calculator. The external anchor test containing 35 items was used to adjust for the slight ability differences between the sample who took the test with a calculator and the sample who took the test without a calculator. The anchor test was administered without a calculator in both groups.

The mean raw score on the test taken with a calculator is 16.36 (SD = 7.99); the mean raw score on the same test for students not allowed to use a calculator is 14.35 (SD = 7.76). The mean raw score on the anchor test for students who took the test with a calculator is 13.77 (SD = 8.27); the mean raw score on the anchor test for students who took the test without a calculator is 13.19 (SD = 8.28). An equipercentile anchor test equating of scores on Test C to scores on Test NC yields the results reported in Table 2. This table shows the difference, in scaled scores, between the conversion for the test given with a calculator and the same test given without a calculator (at selected raw score levels). After equating, the scaled score means on these tests are 450 and 442, a difference of 8 points (this difference in ability between the groups taking the test with and without a calculator is also evident from the raw score differences on the anchor test).[4] The converted scores in Table 2 indicate that the mathematics test given with a calculator is easier than the same test given without a calculator (raw scores convert to lower scaled scores on the test given with a calculator).

[4]There is evidence suggesting that some students in the calculator group used a calculator on the first section, even though that section was intended to be taken without a calculator (the mean on this section is more than .5 raw score points higher in Form C than Form NC). This contamination on the anchor test in terms of calculator use affected the results of Analysis 1, which is why the mean scaled scores are not the same after equating.

Analysis 2

In this analysis, one group of examinees took a test with a calculator and an anchor test with a calculator and another group of examinees took the same test and same anchor test without a calculator. This was accomplished by selecting a subset of 20 items, judged to be unaffected by calculator use, from the 36-item test used in Analysis 1. This subset of items was used as an internal anchor test to adjust for ability differences between the groups taking the test with a calculator and without a calculator. The mean raw score on the test taken with a calculator is 16.36 (SD = 7.99); the mean raw score on the test taken without a calculator is 14.35 (SD = 7.76). The mean raw score on the internal anchor test given with a calculator is 9.10 (SD = 4.93); the mean raw score on the internal anchor test given without a calculator is 7.81 (SD = 4.82). Performance on the total test and anchor test given with a calculator is higher than performance on the same tests given without a calculator.

A linear anchor test equating of scores on the test taken without a calculator to scores on the test taken with a calculator yields the results reported in Table 3. After equating, the conversion for the test given with a calculator is essentially similar to the conversion for a test given without a calculator. The reason for this result is that an advantage due to calculator is similar on the operational test and the anchor test; consequently, score equating is unable to adjust for differences in test difficulty due to calculator use. However, scores on the test given with a calculator are considerably higher than scores on the test given without a calculator. The mean scaled score on the test given with a calculator is 471 while the mean for the test given without a calculator is 442 -- a difference of 29 points.

A Scaling Solution

Because the use of a calculator tends to make SAT mathematical items easier, any equating design that utilizes an anchor test design needs to ensure that the anchor test is administered under the same condition, i.e., with a calculator or without a calculator. In response to this need, an equating study was carried out to ensure that scores on the old SAT-M (administered without a calculator) were comparable to scores on the new SAT-M (administered with a calculator).

The data collection design for the equating study is shown in Table 4. The data collection design for the special administration controlled for calculator use on the anchor test. The equating results from the special administration were compared to equating results based on a data collection design that did not control for calculator use on the anchor test. The results from this special equating study (shown in Table 5) were consistent with those of Study 2, that is, the equating design that adjusted for the effect of calculators resulted in a score conversion that was lower (by ten to twenty scale score points) than the score conversion based on an anchor test equating that did not adjust for the effect of calculators. The results from the special equating study were used for score reporting, thereby maintaining scale comparability between scores on the old and new SAT-M.

Discussion

The two studies described in this paper provide evidence of the existence and magnitude of a context effect due to allowing calculators on a standardized test of mathematical ability. An important question to ask is: what are the implications of this context effect for score reporting?

An important assumption underlying equating models is that the tests to be equated measure the same construct (Linn, 1993; Lord, 1980). According to Linn (1993), "...the goal of estimating the percentage of students scoring above a given level is possible only if the two assessments essentially measure the same thing. That is, it is important that the two assessments be well matched in terms of content coverage, the cognitive demands that are placed on students, and the conditions under which the assessments are administered." This assumption would need to be questioned if data showed that the use of a calculator on the test alters the meaning of the measurements when compared to the test taken without a calculator. Bridgeman et al (1993), and Study 1 in the present paper, provide data on specific items in the SAT that become inappropriate for a test that permits calculators because the skills measured by the item administered with a calculator are quite different when the item is administered without a calculator. For instance, without a calculator a particular item may require estimation skills but with a calculator the estimation skills are no longer needed. These results may imply a shift in the construct being measured by the new test. Research on the dimensional structure of the two tests would be informative in this regard. It would also be useful to assess whether or not the equating relationship between the test given with a calculator and the test given without a calculator is invariant across subgroups (Dorans & Feigenbaum, 1993).

The equating analyses described in Study 2 demonstrate that, under certain circumstances, equating can be used to make scores on a test given with a calculator equivalent to scores on the same test given without a calculator (assuming the two tests are measuring identical constructs). By prohibiting (or allowing) calculator use on the anchor test for both groups, we may assume that the items on the anchor test behave similarly for the group taking the test with a calculator and the group taking the test without a calculator.

Analysis 2 differs from Analysis 1 in assuming that for one group of examinees, calculators would be allowed on both the test being equated and the anchor test. The opposite condition would exist for the other group -- calculators would be prohibited on both the test being equated and the anchor test. Under this circumstance, anchor test equating methods do not work, and scores on the test given with a calculator cannot be made equivalent to scores on the same test given without a calculator. For the same student, scores on the test taken with a calculator will be higher than scores on the test taken without a calculator.

Under the recently approved policy, examinees taking the new SAT-M are permitted to use a calculator on the operational test and the anchor test. In order to provide scaled scores on the 200-to-800 College Board scale, new editions of the SAT are equated to prior forms of the SAT via an anchor test design, where external anchor tests are administered in an unscored section of the test. This situation posed a scaling challenge, because calculators were prohibited on the operational test and the anchor test sections of the old SAT. The scaling study described in the preceding section was the solution to the problem of different calculator conditions for the two testing programs. The evaluation of the results from the scaling study demonstrate that the existence and magnitude of the calculator effect was maintained going from experimental data to data from an operational administration.

It is important to point out that the conclusions from these studies are limited by research designs that did not control for actual use of calculators on the tests. Although calculators were permitted, they were not required, and some students chose not to use a calculator. As a consequence, we don't know the specifics on calculator use with respect to individual items or sets of items, information that could inform calculator use policies. Nevertheless, data from the special equating study described above indicate that 89% of the test takers

brought a calculator to the test. Furthermore, among the 89% who brought calculators to the test, 62% reported using the calculator on a few of the test questions, 19% reported using the calculator on about a third of the questions, and 11% reported using the calculator on about half of the questions. These percentages suggest that most students were using calculators when permitted to do so, and that they were using the tool thoughtfully. Stronger inferences, especially at the item and item type level, awaits the collection of more detailed data.

References

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service.

Bridgeman, B., Harvey, A., & Braswell, J. (1993). *Effects of calculator use on SAT-M performance.* Paper presented at the annual meeting of AERA, Atlanta.

Cohen, A. S., & Kim, S. (1992). Detecting calculator effects on item performance. *Applied Measurement in Education, 5,* 303-320.

Dorans, N. J. & Feigenbaum, M. D. (1993). *Equating issues engendered by changes to the SAT and PSAT/NMSQT.* Paper presented at the annual meeting of NCME, Atlanta.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83-102.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Morgan, R., & Stevens, J. (1991). *Experimental study of the effects of calculator use on the Advanced Placement Calculus examinations* (Research Report No. RR-91-5). Princeton, NJ: Educational Testing Service.

Rigol, G. (1993). *Calculators: Balancing educational, administrative, and equity issues.* Paper presented at the annual meeting of NCME, Atlanta.

Figure 1

Items That Are Easier With A Calculator (Study 1)

For each question the examinee is asked to compare the quantity in column A to the quantity in column B and indicate

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

	Column A	Column B
Item 17	The number of seconds in 24 hours	The number of minutes in eight weeks
Item 19	32/33	42/43
Item 22	$1{,}000 \times 0.05$	$\frac{1}{0.2 \times 0.05}$

Key: Item 17 -- A

Item 19 -- B

Item 22 -- B

Table 1

Difficulty Estimates (Percent Correct) from the Calculator and No Calculator Groups

	Testing Condition		
Item	Calculator	No Calculator	Difference
1	.92	.91	.01
2	.94	.89	.05
3	.78	.77	.01
4	.83	.81	.02
5	.85	.85	.00
6	.68	.68	.00
7	.72	.72	-.00
8	.62	.62	.00
9	.81	.80	.01
10	.69	.69	.00
11	.71	.67	.04
12	.73	.73	.00
13	.72	.71	.01
14	.58	.58	.00
15	.56	.57	-.01
16	.57	.56	.01
17	.62	.51	.11
18	.42	.40	.02
19	.69	.50	.19
20	.43	.44	-.01
21	.39	.39	.00
22	.57	.32	.25
23	.17	.13	.04
24	.24	.23	.01
25	.15	.15	.00

Table 2

Equating Results Controlling for Calculator Effect on

Anchor Test Scores

	Scaled Score			Cumulative Frequency	
Raw Score	Calculator	No Calculator	Scaled Score Difference	Calculator	No Calculator
36	770	770	0	1900	1860
35	760	750	10	1897	1859
30	660	680	-20	1812	1820
25	580	610	-30	1625	1699
20	500	530	-30	1316	1446
15	430	450	-20	931	1061
10	350	370	-20	477	630
5	290	300	-10	157	256
0	230	240	-10	18	32
Mean	450	442			
SD	118	118			

Table 3

Equating Results NOT Controlling for Calculator
Effect on Anchor Test Scores

	Scaled Score			Cumulative Frequency	
Raw Score	Calculator	No Calculator	Scaled Score Difference	Calculator	No Calculator
36	770	770	0	1900	1860
35	750	750	0	1897	1859
30	680	680	0	1812	1820
25	610	610	0	1625	1699
20	530	530	0	1316	1446
15	450	450	0	931	1061
10	370	370	0	477	630
5	300	300	0	157	256
0	250	240	10	18	32
Mean	471	442			
SD	122	118			

Table 4
Data Collection Design for Equating Study to
Adjust for Calculator Effect

Data Source	Sample	Anchor Test	Total Test
Special Administration	Old Form	No Calculator	No Calculator
Special Administration	New Form	No Calculator	Calculator
National Administration	Old Form	No Calculator	No Calculator
National Administration	New Form	Calculator	Calculator

Table 5
Results of Equating Study to
Adjust for Calculator Effect

Raw Score	Scaled Score NOT Adjusting for Calculator Effect	Scaled Score Adjusting for Calculator Effect	Scaled Score Difference
60	770	780	-10
55	720	730	-10
50	670	670	0
45	630	620	10
40	580	580	0
35	550	530	20
30	510	490	20
25	470	450	20
20	430	410	20
15	390	380	10
10	350	340	10
5	300	290	20
0	260	250	10
Mean	595	484	
SD	121	124	