J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 5, Number 7 · March 2007

The Effect of Using Item Parameters Calibrated from Paper Administrations in Computer Adaptive Test Administrations

Mary Pommerich

www.jtla.org

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

Volume 5, Number 7

The Effect of Using Item Parameters Calibrated from Paper Administrations in Computer Adaptive Test Administrations

Mary Pommerich

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:

Pommerich, M. (2007). The effect of using item parameters calibrated from paper administrations in computer adaptive test administrations. *Journal of Technology, Learning, and Assessment, 5*(7). Retrieved [date] from http://www.jtla.org.

J·T·L·A

Abstract:

Computer administered tests are becoming increasingly prevalent as computer technology becomes more readily available on a large scale. For testing programs that utilize both computer and paper administrations, mode effects are problematic in that they can result in examinee scores that are artificially inflated or deflated. As such, researchers have engaged in extensive studies of whether scores differ across paper and computer presentations of the same tests. The research generally seems to indicate that the more complicated it is to present or take a test on computer, the greater the possibility of mode effects. In a computer adaptive test, mode effects may be a particular concern if items are calibrated using item responses obtained from one administration mode (i.e., paper), and those parameters are then used operationally in a different administration mode (i.e., computer). This paper studies the suitability of using parameters calibrated from a paper administration for item selection and scoring in a computer adaptive administration, for two tests with lengthy passages that required navigation in the computer administration. The results showed that the use of paper calibrated parameters versus computer calibrated parameters in computer adaptive administrations had small to moderate effects on the reliability of examinee scores, at fairly short test lengths. This effect was generally diminished for longer test lengths. However, the results suggest that in some cases, some loss in reliability might be inevitable if paper-calibrated parameters are used in computer adaptive administrations.

J·T·L·A

The Effect of Using Item Parameters Calibrated from Paper Administrations in Computer Adaptive Test Administrations

Mary Pommerich

Introduction

As computer technology becomes more prevalent and accessible, we are likely to see more testing programs shifting from paper administration to computer administration. Bennett (2002) argued that the use of computer technology in testing is inevitable, because computer technology is being increasingly used for instruction. This contention is supported by reports of high levels of computer use in schools and at home. For example, in 2001, the U.S. Department of Commerce reported computer use at school by 74% of 5–9 year olds and 85% of 10–17 year olds (2002). In 2003, the U.S. Department of Education reported that about 91% of children in nursery school through grade 12 use computers, including 97% of 9th–12th graders (Debell & Chapman, 2006). These numbers indicate that computers are becoming increasingly common in our society.

For testing programs that convert from paper to computer administration, some may administer computer adaptive tests (CATs), while others may continue to administer fixed form tests, but do so via computer. Some programs may continue to test under both paper and computer platforms, while other programs may drop paper administrations altogether. Programs that utilize computer administration may be likely to take further advantage of computer capabilities, and compute scores based on Item Response Theory (IRT). IRT is also used in CAT administrations for selecting items for administration and for scoring responses.

The use of IRT for item selection and/or scoring necessitates calibrating item parameters. For tests that are administered on the computer, it is plausible that for convenience purposes, data from paper administrations might be used to calibrate item parameters for initial use with computer administration. Paper administration is a much quicker and more efficient means of collecting sufficient item response data for calibrating than is computer administration. For example, when starting a computerized adaptive testing program, it might be very costly and time-consuming to calibrate the initial pool(s) using data from computer administrations of the items. Every item in the pool would have to be administered to a sufficient number of examinees in order to calibrate. If the item pool is large, this would require a substantial amount of testing that likely could not be done quickly (or cheaply) via computer administration, and could result in unwanted exposure of the item pool even before it becomes operational. Pre-testing via paper administration could alleviate some of these costs and security concerns.

Thus, a testing program might consider initially using item parameters calibrated from paper and pencil administrations of the items for operational computer administrations, until enough data have been collected to calibrate from computer administrations. Wang and Kolen (2001) noted that in practice, item parameters from paper calibrations are often used in computer adaptive administrations. However, Kolen (1999–2000) cautioned against assuming that items behave in the same way across paper-and-pencil and computer adaptive tests. Likewise, Parshall, Spray, Kalohn, and Davey (2002) cautioned that item calibrations based on paper and pencil administrations might not represent the performance of those same items in a computer administration. If items are calibrated using item responses from one medium and then used operationally in another medium, examinees could be adversely affected in cases where there are parameter differences across modes of administration.

An abundance of research has been conducted evaluating the comparability of computer administered and paper administered tests. In a recent review of trends in comparability studies, Paek (2005) asserted that sufficient evidence exists to conclude that computer administration does not significantly affect student performance, with the exception of tests containing lengthy reading passages. The comparability research generally seems to suggest, however, that mode differences might be influenced by the degree to which the presentation of the test and the process of taking the test differ across the modes of administration (Pommerich, 2004). Whether scrolling or some other form of navigation is necessary in the computer mode appears to be a key component to the likelihood of mode effects.

Conflicting results have been found across studies about the suitability of using paper calibrated parameters in a computer administration. Hetter, Segall, and Bloxom (1997) concluded that paper calibrated parameters could be used in a computer adaptive test without changing the construct being measured or reducing reliability. Choi and Tinkler (2002) found that for a third grade reading test, scores for computer examinees would be substantially lower if paper-based calibrations were used in scoring rather than computer-based calibrations, and concluded that computer scores might be different if paper calibrated parameters were used in a computer administration. It is important to note that the Hetter et al. conclusion was in regard to tests containing items that could all be displayed in their entirety on a single screen, whereas the Choi and Tinkler conclusion was in regard to a test that required scrolling administered in a population that did not have much experience testing on computers (third graders).

This study evaluates differences in item parameters across calibrations conducted from paper and computer administrations of the same fixed form tests, for passage-based tests in the content areas of Reading and Science Reasoning. More specifically, this study addresses the question of whether parameter differences observed across administration mode are practically significant by examining the effect of using the different parameters in computer adaptive administrations. Item parameters may differ across paper and computer calibration samples, but if the differences are not of a magnitude to adversely affect an examinee's score if he or she were to take the items in one mode versus another, then we may not need to be concerned about the differences. The effects of using item parameters calibrated from paper and computer administrations in a CAT are studied via a simulation where each set of parameters is used to select items and score responses. The simulation study is based on real parameter differences that were observed across paper and computer administrations of the tests.

Data Source

The paper and computer administrations occurred as part of a large-scale comparability study that was conducted between October 2000 and January 2001.[1] A total of 61 schools participated in the study, with 11th and 12th grade students testing. The 61 schools were part of a nation-wide random sample of 720 schools that were solicited to participate in the study. The participating schools agreed to test juniors and seniors that had taken or planned to take an elective national standardized achievement test for college applicants. Testing was conducted at a classroom level, with schools selecting classes to participate in the study.

Fixed form tests were administered in the content areas of English, Math, Reading, and Science Reasoning, with the same test administered across modes. The test forms administered for Reading and Science Reasoning were intact forms from a national testing program that had previously been administered operationally via paper and pencil. The test forms administered for English and Math contained a representative subset of items selected from operational forms to accommodate a shorter testing period than used operationally. Because only the Reading and Science Reasoning forms met the specifications for paper forms, only results for Reading and Science Reasoning were studied here.

Within a classroom, examinees were randomly assigned to a paper or computer administration. Approximately one third of each class took the paper-and-pencil test, with approximately two thirds taking the computerized test. Examinees assigned to the paper mode were randomly assigned to a content area (English, Math, Reading, or Science Reasoning). Examinees assigned to the computer mode were randomly assigned to a content area and navigation variation (English Navigation 1, Math, Reading Navigation 1, Science Reasoning Navigation 1, English Navigation 2, Reading Navigation 2, or Science Reasoning Navigation 2). The navigation variations studied for Reading and Science Reasoning were paging and scrolling. The Reading and Science Reasoning tests were administered across all modes and navigation variations with the same time constraints as used operationally (35 minutes).

Booklet Versus Computer Presentation

The Reading test contained four passages with 10 multiple-choice items in each passage (40 items total). The Science Reasoning test contained seven passages with five to seven multiple-choice items in each passage (40 items total). For both tests, the length of the passages meant that some form of navigation was necessary in the computer administration to view the entire passage. By necessity, then, there were substantial presentation differences across paper and computer administrations of the tests.

Booklet Presentation

In the booklet presentation of the Reading and Science Reasoning tests, each passage was presented first in its entirety, followed by the test items. The Reading passages were presented in two columns per page. For both the Reading and Science Reasoning tests, each passage and accompanying items occupied two booklet pages, so that all information for a passage could be viewed at once over a two-page spread. Examinees were able to move freely throughout the passages and/or items in the booklet while taking the test. They could respond to items and passages in any order, and were not required to give responses to all items.

Computer Presentation

In the computer presentation, a 17-inch monitor was used, set to a resolution of 1280 × 1024. For both the Reading and Science Reasoning tests, passages and items were presented jointly on the screen, with the passage appearing in a window on the left half and an individual item appearing in a window on the right half of the screen. Only a portion of the passage was visible on screen at once. Thus, examinees had to use some form of navigation to read the entire passage (either scrolling or paging). In the scrolling variation, examinees could scroll line-by-line, or manipulate a sliding scroll bar to move quickly through the passage. In the paging variation, the passage was divided into separate pages and the examinee moved between pages by clicking on a specific page number, or by using “Next Page” or “Previous Page” buttons. Examinees moved between items by clicking on a specific item number or by using “Next Question” or “Previous Question” buttons. Within a passage, examinees were allowed to answer items in any order. They were required to answer all items prior to moving on to the next passage. Once an examinee completed a passage and moved on to the next passage, they were not allowed to return to the previous passage. Also, passages were presented one at a time, so that examinees could not see the next passage until they proceeded to it. All computer examinees took a short tutorial prior to testing that demonstrated how to use all of the functions necessary to take the computerized test.

Results

Analysis Samples

Due to noted irregularities during assignment to a testing condition or during testing, a small percentage of records (~5%) were deemed unusable; records that were problematic were deleted from the final analyses. There appeared to be no systematic pattern to the data loss. The resulting sample sizes for the analyses are reported in Table 1.[2] To check the assumption of random equivalence of the groups, χ^2 tests of independence were conducted for Reading and Science Reasoning to evaluate the relationship between analysis group (Paper, Computer Page, Computer Scroll) and available demographic variables (gender, race/ethnicity, grade, and plans to attend college). For both Reading and Science Reasoning, the results for each demographic variable were non-significant, which suggested that the distributions of the demographic variables were similar across the analysis groups. The results of the χ^2 tests suggest that the groups can be considered randomly equivalent.

Table 1: Final Sample Sizes for Analyses

Test	Mode	Condition	N
Reading	Computer	Page	996
	Computer	Scroll	1089
	Paper	—	1086
Science Reasoning	Computer	Page	902
	Computer	Scroll	1067
	Paper	—	1055

Within school sample sizes for the analysis sample ranged from 49 to 279 students, with a median sample size of 98 students. The gender breakdown for the analysis sample was 56% female and 44% male. The race/ethnicity breakdown was 70% Caucasian, 15% African-American, 3% Mexican-American or other Hispanic, 2% Asian-American, 1% American-Indian, 1% Multiracial, and 8% reporting "Other" or opting not to respond. The grade breakdown was 63% 11th graders and 36% 12th graders. 89% of the analysis sample had plans to attend college. For study participants that had already taken the elective national standardized achievement test for college applicants (~55% of the analysis sample), their average scores on that test were 21.5 in Reading and 21.2 in Science Reasoning. Average scores for the national population of examinees in the corresponding test year were 21.3 in Reading and 21.0 in Science Reasoning. These results suggest that the academic ability of the analysis sample was likely similar to the academic ability of the population taking the elective national standardized achievement test for college applicants.

Total Score Performance

Table 2 gives the difference in average total scores across modes for the two computer conditions (computer – paper), and the value of the t-statistic for a test of the hypothesis that the average scores are equal across paper and computer modes. Positive values indicate a higher average score on computer than on paper. Table 2 shows a significant difference in average scores only for the Science Page condition, which favored computer examinees. Average scores for the Science Scroll condition also favored computer examinees, but not significantly. Average scores for both of the Reading conditions favored paper examinees, but the score difference was not significant.

Table 2: Difference in Average Total Scores Across Modes (Computer – Paper), and t-Statistic for Comparison of the Average Scores

Test	Computer Condition	Difference	T
Reading	Page	–0.21	–0.66
	Scroll	–0.25	–0.82
Science Reasoning	Page	+0.73	+2.41*
	Scroll	+0.44	+1.50

*p <. 05

The non-significant results for total score performance for both Reading computer conditions suggest that it might be possible to combine data from the Page and Scroll conditions for subsequent analyses. However, item level analyses of both Reading and Science Reasoning showed that examinees responded differently to some items under the Page and Scroll conditions and that more items favored computer examinees in the Page condition than in the Scroll condition (Pommerich, 2004). Because CATs are highly dependent on item-level parameters and there was evidence of some degree of differential item-level performance across the two navigation variations, separate analyses were conducted for the Page and Scroll conditions, for both Reading and Science Reasoning.

Item Parameter Differences

Item responses for the paper and computer samples were calibrated under a three-parameter logistic (3PL) model using Bilog-MG. The calibrations were conducted under the same conditions across the paper and computer samples. Not-reached and omitted responses were treated as incorrect because that is how they are scored operationally for paper examinees in the national testing program. Correlations between paper and computer item parameters for the Reading and Science Reasoning tests are given in Table 3. The *b* parameters were very highly correlated. The *a* parameters were less highly correlated than the *b* parameters, but were still highly correlated. The *c* parameters, which are typically less-well estimated than the *a* and *b* parameters, were moderately correlated across modes.

Table 3: Correlation Between Computer and Paper Item Parameters

Test	Computer Condition	Correlation		
		a	*b*	*c*
Reading	Page	0.82	0.96	0.65
	Scroll	0.77	0.93	0.59
Science Reasoning	Page	0.68	0.93	0.45
	Scroll	0.80	0.95	0.78

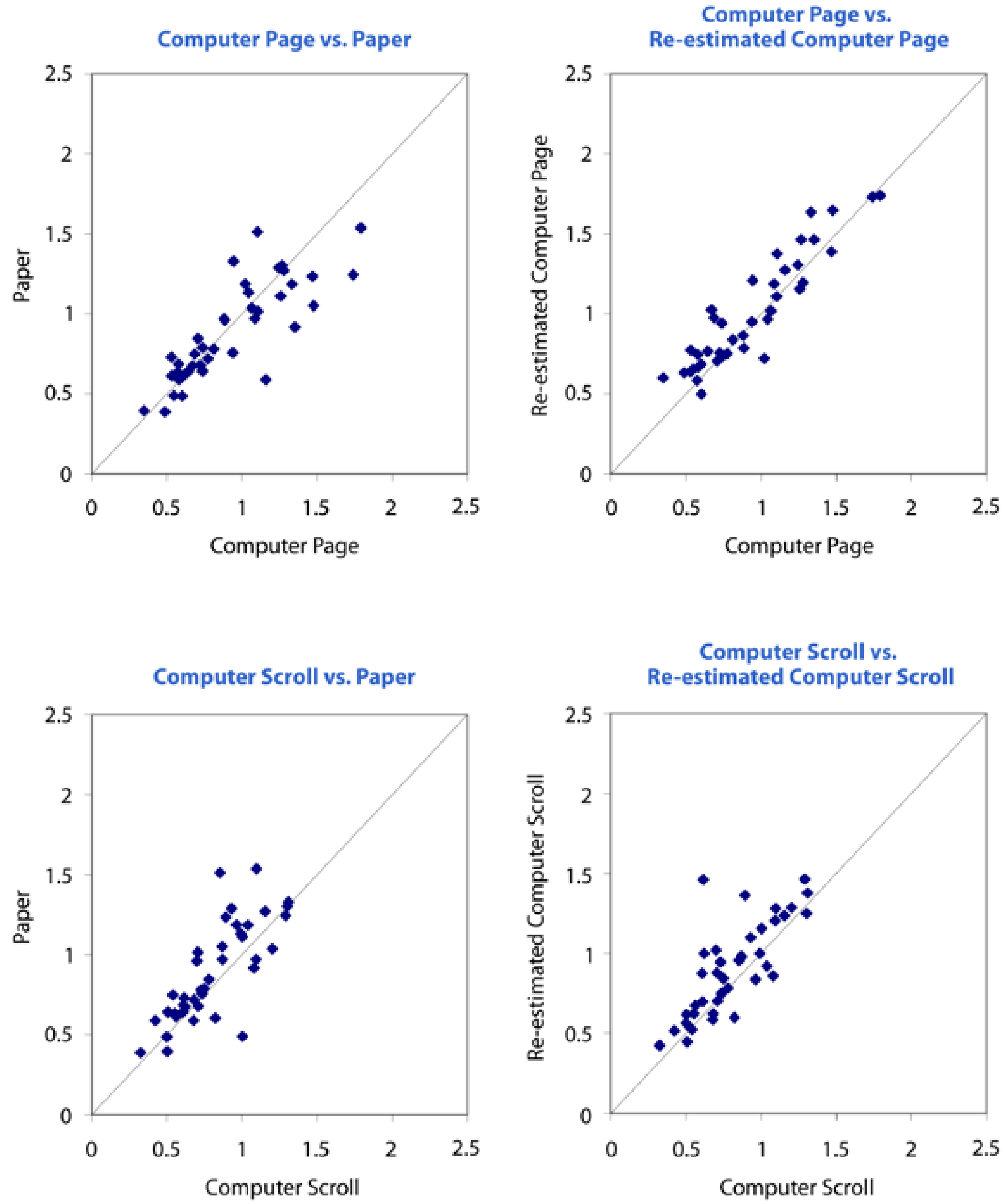
Correlations between the computer parameters and "re-estimated" computer parameters are given in Table 4 (next page), as a baseline com–parison for the correlations between the computer and paper parameters. To compute the re-estimated computer parameters, the 3PL computer parameters were used to generate item responses in a normally distributed sample of examinees of the same size as the original calibration sample. The simulated 0,1 responses were then calibrated (i.e., the "true" computer parameters were re-estimated) under the same conditions that they were originally calibrated. A comparison of the original (true) parameters with the re-estimated parameters provides a baseline measure of how much difference we might expect to see between the parameters simply due to estimation error in the calibration process. The correlations reported in Table 4 show there was some estimation error associated with the calibration process, mainly in the *a* and *c* parameters. If there were no mode effects, we would expect to see correlations of a similar magnitude between the paper and computer parameters. However, a comparison of Tables 3 and 4 shows higher correlations between the computer and re-estimated computer parameters than between the computer and paper parameters, which suggests that there were some mode effects contributing to the observed paper and computer parameter differences, above and beyond calibration error.

Table 4: Correlation Between Computer and Re-estimated Computer Item Parameters

Test	Computer Condition	Correlation		
		a	***b***	***c***
Reading	Page	0.92	0.99	0.76
	Scroll	0.78	0.99	0.82
Science Reasoning	Page	0.86	0.99	0.84
	Scroll	0.90	0.99	0.82

Figure 1 shows comparison plots of the a parameters for Reading, one each for the computer page parameters versus the paper parameters (top left), the computer page parameters versus the re-estimated computer page parameters (top right), the computer scroll parameters versus the paper parameters (bottom left), and the computer scroll parameters versus the re-estimated computer scroll parameters (bottom right).

Figure 1: Comparison of *a* Parameters for Reading



Computer Page vs. Paper

Computer Page vs. Re-estimated Computer Page

Computer Scroll vs. Paper

Computer Scroll vs. Re-estimated Computer Scroll

Figures 2–3 show similar plots for the Reading *b* and *c* parameters.

Figure 2: Comparison of *b* Parameters for Reading

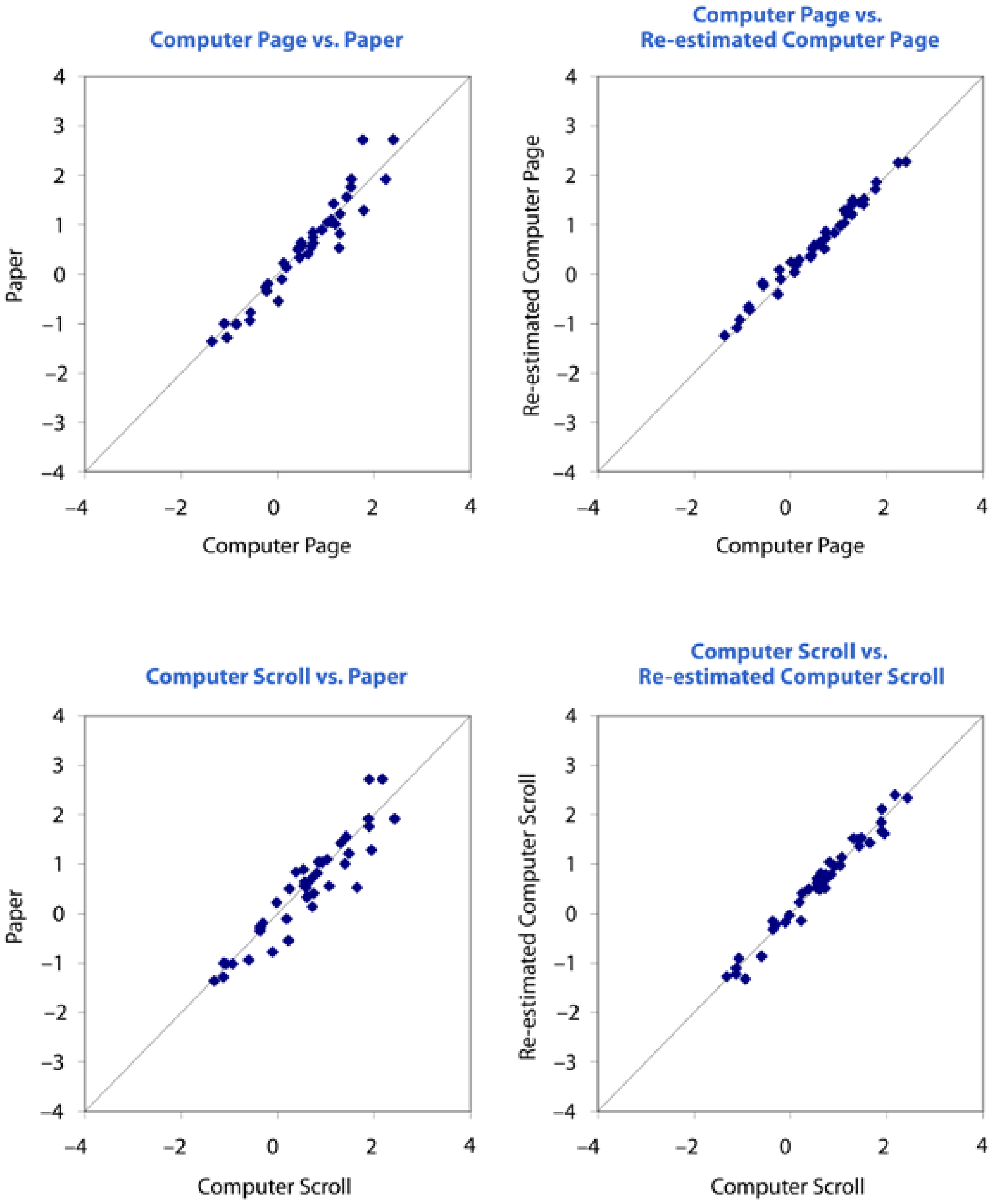
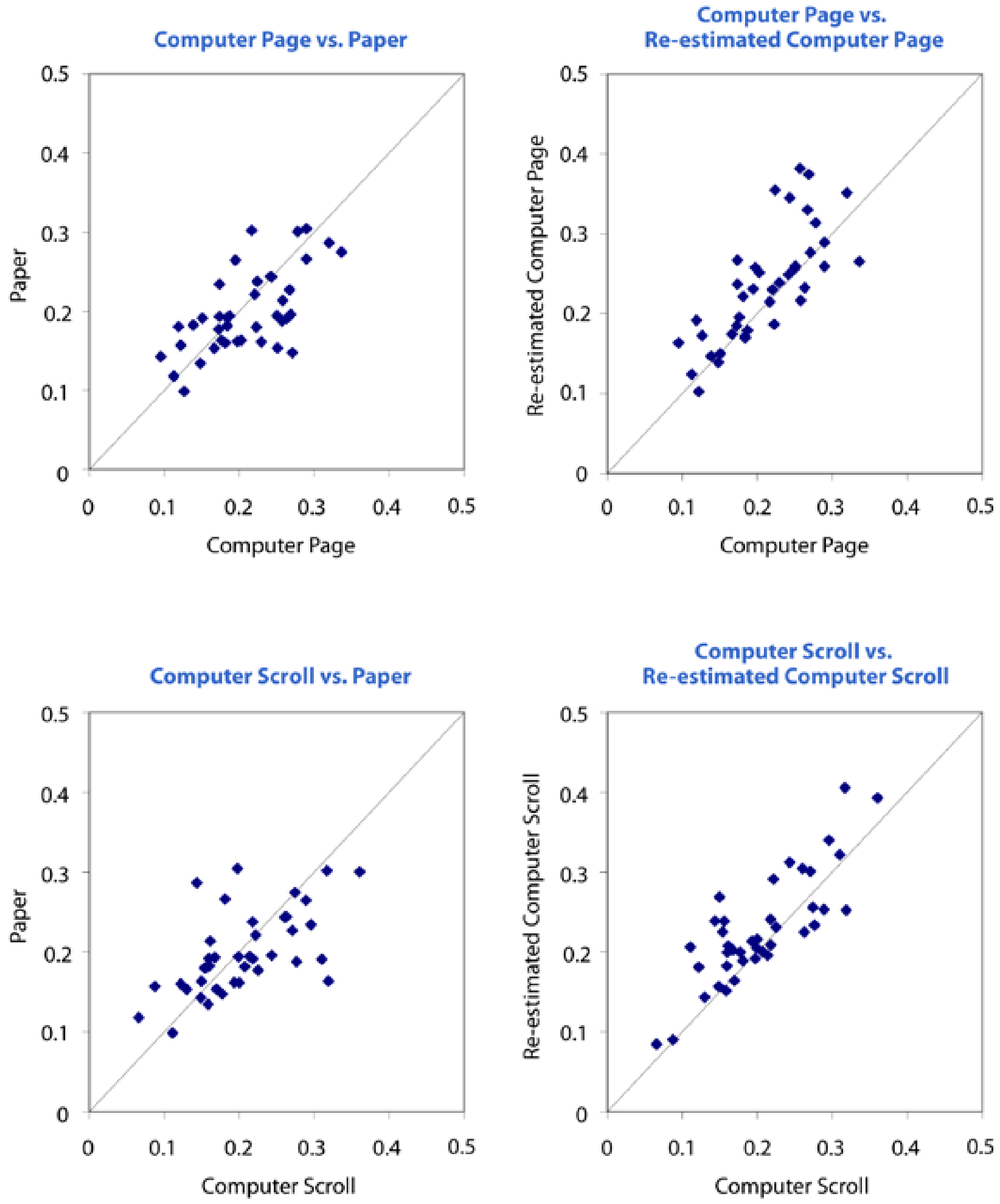


Figure 3: Comparison of *c* Parameters for Reading



Computer Page vs. Paper

Computer Page vs. Re-estimated Computer Page

Computer Scroll vs. Paper

Computer Scroll vs. Re-estimated Computer Scroll

Likewise, Figures 4–6 show similar plots for the Science Reasoning *a*, *b*, and *c* parameters.

Figure 4: Comparison of *a* Parameters for Science Reasoning

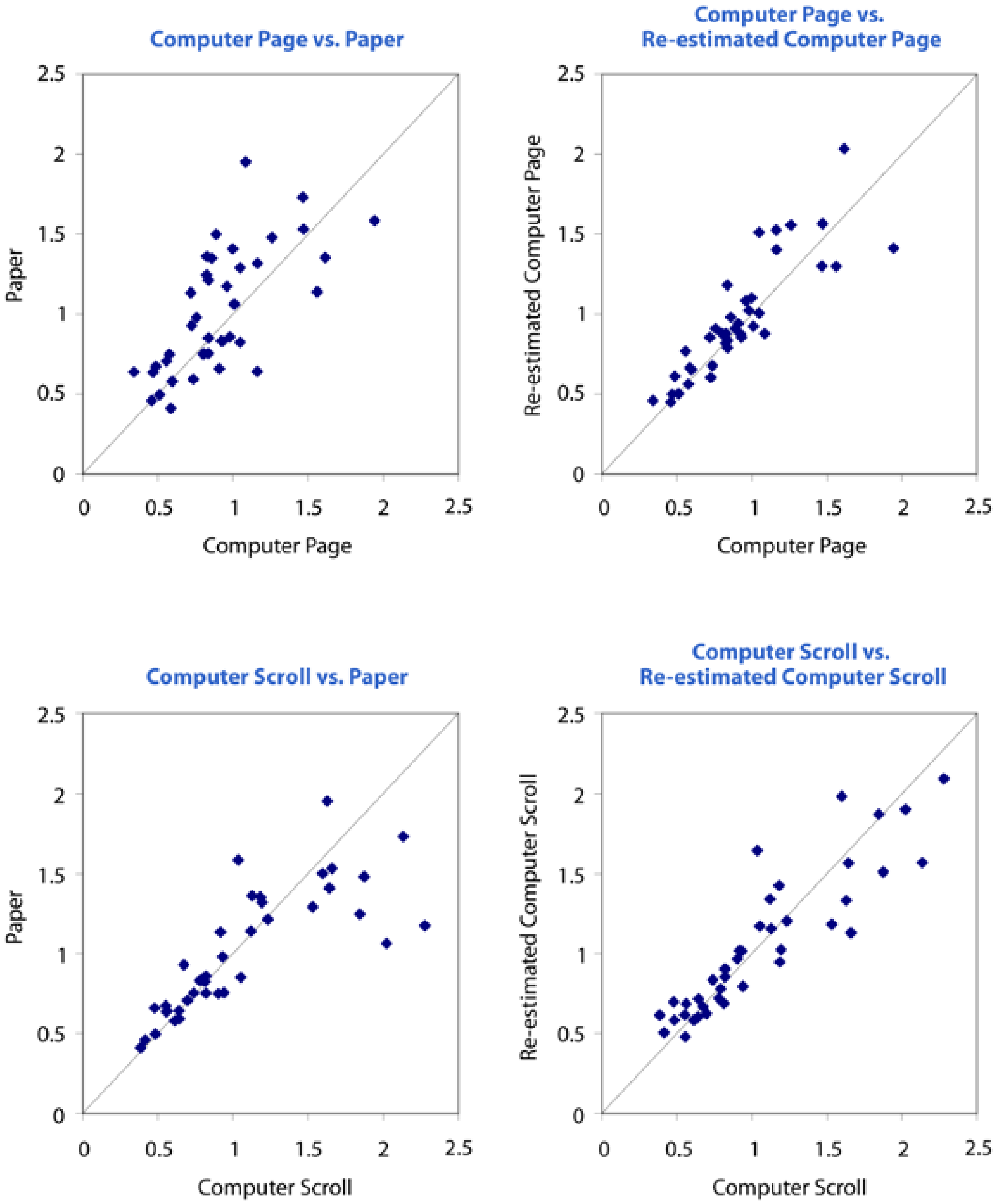
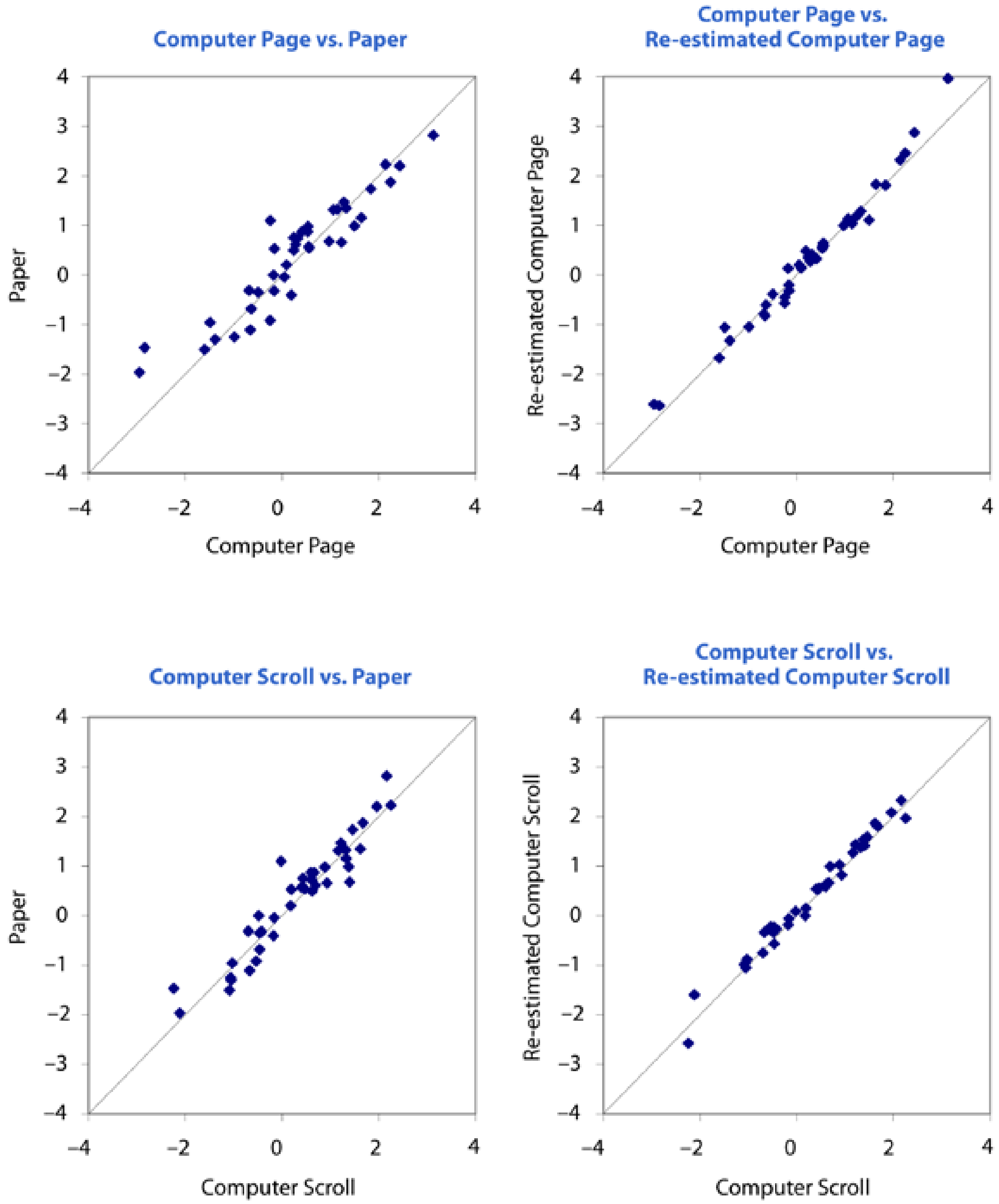


Figure 5: Comparison of *b* Parameters for Science Reasoning



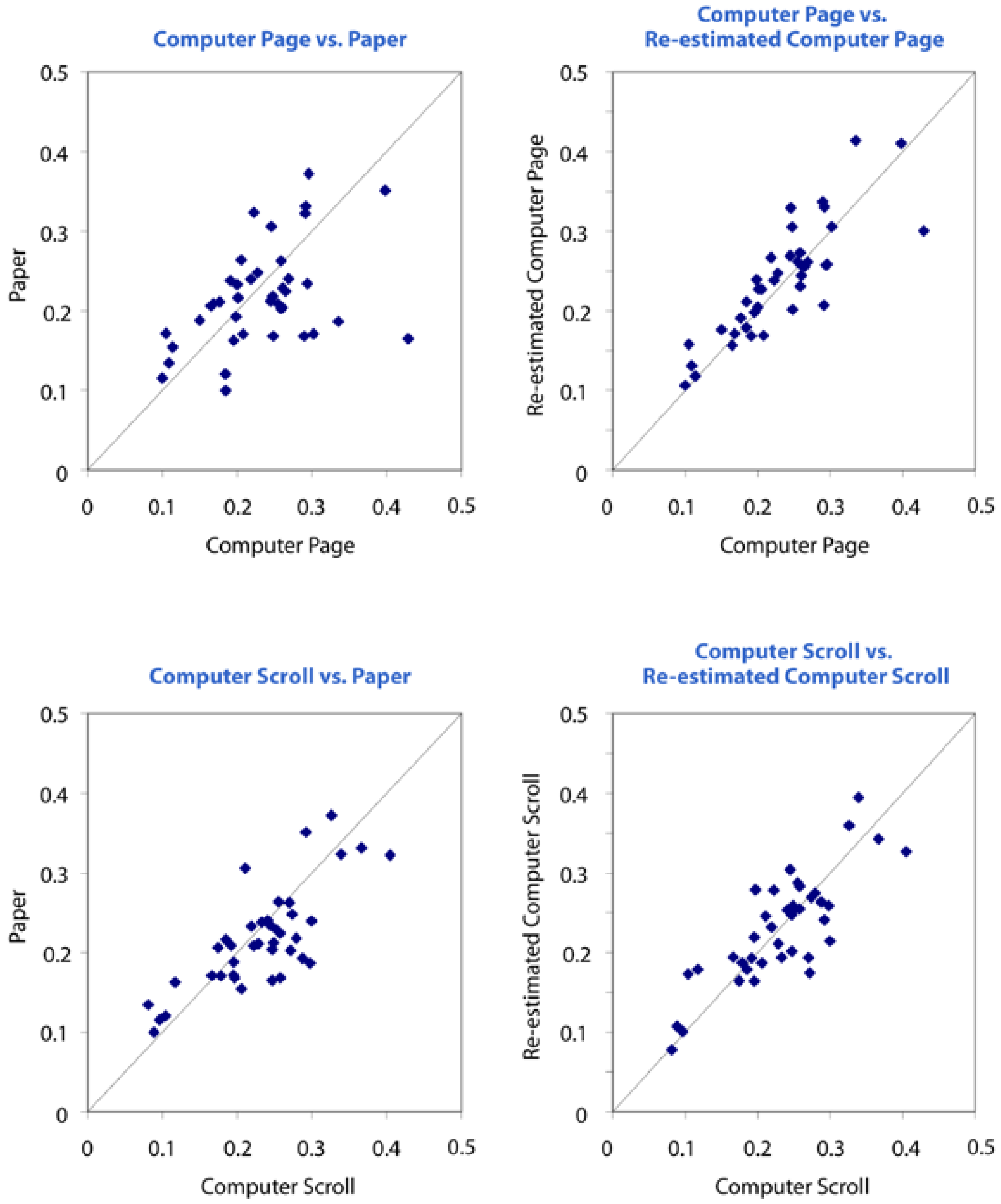
Computer Page vs. Paper

Computer Page vs. Re-estimated Computer Page

Computer Scroll vs. Paper

Computer Scroll vs. Re-estimated Computer Scroll

Figure 6: Comparison of *c* Parameters for Science Reasoning



Across Figures 1–6, plots of the computer versus paper parameters consistently showed more spread than the plots of the computer versus re-estimated computer parameters, for both the Page and Scroll conditions, again suggesting that there were some mode effects contributing to the observed parameter differences.

Are the Parameter Differences Important?

Table 2 (page 10) showed us that administration mode significantly affected average total scores only in the case of Science Reasoning Page. However, a comparison of item-level performance on Reading and Science Reasoning showed significant performance differences across the paper and computer administration modes for some items, for both the paging and scrolling conditions (Pommerich, 2004). It is likely that the parameter differences observed across the computer and paper calibration samples are attributable to some degree to these item-level mode effects. But whether these parameter differences are of any practical significance is difficult to gauge without further analysis. One way to quantify the effect of the parameter differences is to evaluate how examinee scores are affected when item parameters calibrated from a paper administration are used in operational CAT administrations. If examinee CAT scores are not adversely affected by the use of paper-calibrated parameters, then we likely need not be concerned about the mode effects.

To address the importance of the observed parameter differences, a CAT was simulated under different conditions and test-retest reliability was compared across the conditions. In order to create CAT pools, each set of parameters was cloned to create item pools eight times the size of the original form (i.e., the item parameters were repeated eight times). This simulation assumed that the item parameters from one fixed form were representative of what the item parameters would be in an operational pool built from unique items. Thetas for 10,000 examinees were generated from a N(0,1) distribution and each examinee was administered two adaptive tests using maximum information item selection and exposure control parameters computed using the Sympson-Hetter algorithm (Sympson & Hetter, 1985; Hetter & Sympson, 1997). The Pearson product-moment correlation was computed between the final ability estimates from the two tests, as a measure of test-retest reliability. Simulations were conducted for Reading and Science Reasoning under both the scroll and page conditions, for fixed test lengths ranging from 10 to 40 items.

Table 5 (next page) summarizes the three conditions that were simulated for each content area and navigation variation, labeled "True," "EstC," and "EstP." The three conditions differed in terms of the item parameters used to generate responses, select items, and score responses.

Table 5: **Parameters Used to Generate Responses, Select Items, and Score Responses for the Three Simulation Conditions**

Condition	Generate Responses	Select Items	Score Responses
True	Computer	Computer	Computer
EstC	Computer	Re-estimated Computer	Re-estimated Computer
EstP	Computer	Paper	Paper

Under the condition labeled "True," the computer parameters were treated as the true parameters, and were used to generate item responses, select items, and score responses. Specifically, items were selected based on information tables and exposure control parameters computed from the computer parameters, responses were generated based on the computer parameters, and intermediate and final ability estimates were based on the computer parameters. This condition represents what would happen operationally if true computer parameters were used in a computer adaptive administration. It represents the ideal case for test-retest reliability under the computer parameters.

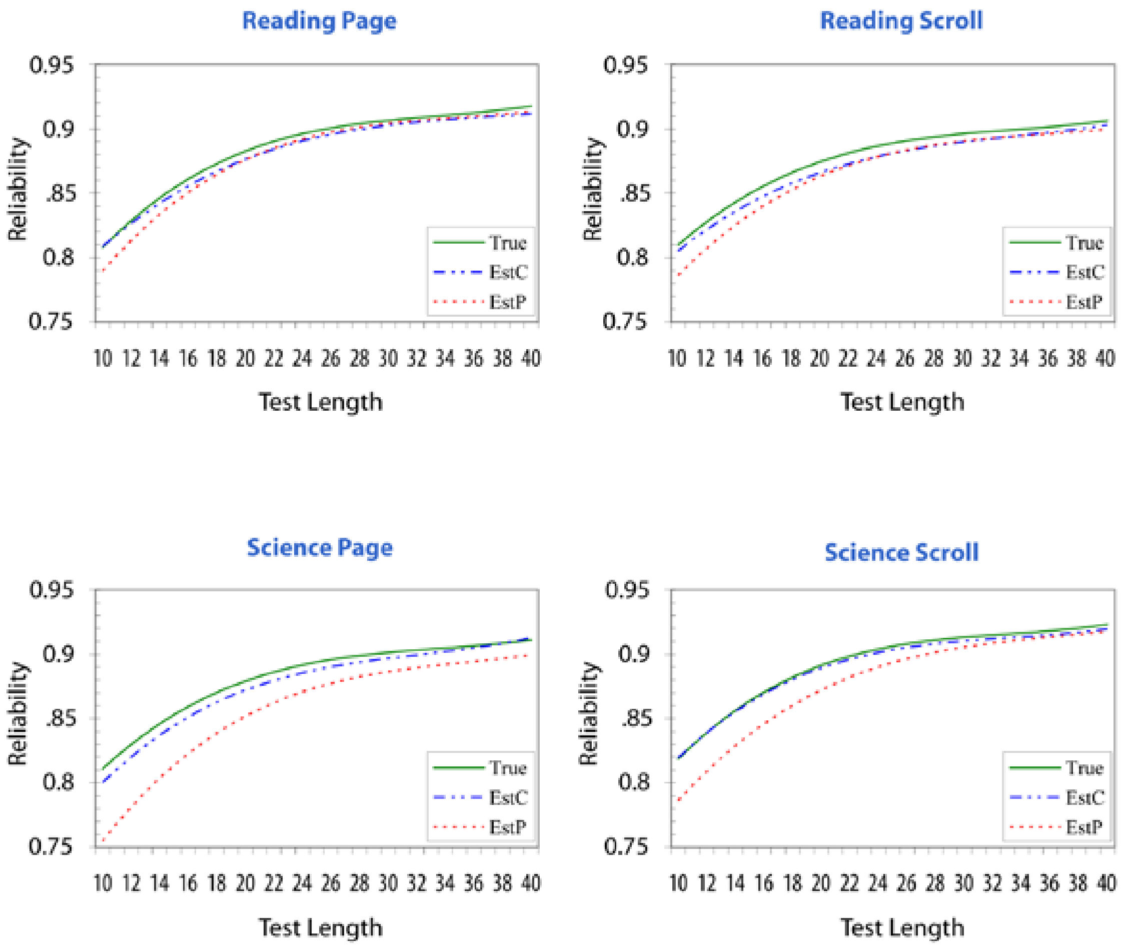
Under the condition labeled "EstC," the computer (true) parameters were used to generate item responses and the re-estimated computer parameters were used to select items and score responses. Specifically, items were selected based on information tables and exposure control parameters computed from the re-estimated computer parameters, responses were generated based on the computer (true) parameters, and intermediate and final ability estimates were based on the re-estimated computer parameters. This condition represents what would happen operationally if estimated computer parameters were used in a computer adaptive administration (as opposed to true computer parameters). Since the true computer parameters would never be known in practice, this condition represents the realistic case for test-retest reliability under the computer parameters.

Under the condition labeled "EstP," the computer (true) parameters were used to generate item responses and the paper parameters were used to select items and score responses. Specifically, items were selected based on information tables and exposure control parameters computed from the paper parameters, responses were generated based on the computer parameters, and intermediate and final ability estimates were based on the paper parameters. This condition represents what would happen operationally if the calibrated paper parameters were used in a computer adaptive administration.

A comparison of the results for the True and EstC conditions should indicate how much test-retest reliability is affected by estimation error in the calibration process to obtain the computer parameters. A comparison of the results for the EstC and EstP conditions should indicate how much test-retest reliability is affected by the use of paper parameters in a computer adaptive administration.

The simulated test-retest reliabilities were modeled using a third degree polynomial regression model, with test length as the predictor. This model accounted for approximately 99% of the variance in the test-retest reliabilities. Figure 7 (next page) shows plots of the predicted test-retest reliabilities by test length for the True, EstC, and EstP conditions for Reading Page (top left), Reading Scroll (top right), Science Page (bottom left), and Science Scroll (bottom right), respectively. The results for both Reading and Science Reasoning show some loss in reliability simply due to calibrating the item parameters used in the item selection and scoring (i.e., reliability is lower for the EstC condition than the True condition). The results also suggest that we can expect some loss in reliability above and beyond the loss due to calibrating the computer item parameters, if we were to use paper-calibrated parameters in a computer adaptive administration (i.e., reliability is lower for the EstP condition than the EstC condition).

Figure 7: Predicted Test-Retest Reliabilities for the Simulated Conditions



The loss in reliability due to the use of paper parameters was greater for Science Reasoning than for Reading, which corresponds to both the larger differences in average total score that were observed for Science Reasoning (Table 2, page 10), and the slightly greater spread observed for Science Reasoning in the plots of the computer versus paper parameters (Figures 1–6, pages 13–18). The loss in reliability due to the use of paper parameters in a computer administration was greatest for test lengths of 10 for all conditions. As test length increased, the reliability observed when the paper parameters were used in the CAT administration approached the reliability observed when the re-estimated computer parameters were used in the CAT administration. This was true for all conditions but Science Page (which showed a significantly higher average total score for computer examinees than for paper examinees).

The simulation results imply that operationally, if the CAT test length is set to meet a target reliability, then a slightly longer test may be required if paper calibrated parameters are used rather than the computer calibrated parameters in the CAT administration. To determine how much longer test length would need to be, another simulation was conducted. In the new simulation, target reliabilities were computed for each computer condition as the correlation between test-retest number right scores on a fixed-form test for 10,000 normally distributed examinees. Item responses were generated for the fixed form using the computer (true) parameters. The test-retest reliability from this simulation represents the reliability of the fixed form in the calibration sample. The target reliabilities were .85, .84, .84, and .85 for Reading Page, Reading Scroll, Science Reasoning Page, and Science Reasoning Scroll, respectively. Results from the polynomial model for the EstC condition (which represents the realistic case for test-retest reliability under the computer parameters) show that the target reliabilities are met with CAT test lengths of 15, 15, 14, and 13 for Reading Page, Reading Scroll, Science Reasoning Page, and Science Reasoning Scroll, respectively.

In order to meet the target reliabilities for Reading Page and Reading Scroll, one additional item is needed if the paper parameters are used rather than the re-estimated computer parameters in the CAT administration. This corresponds to a test length that is approximately 7% longer. For Science Reasoning Scroll, three additional items are needed if the paper parameters are used rather than the re-estimated computer parameters in the CAT administration. This corresponds to a test length that is approximately 23% longer. For Science Reasoning Page, four additional items are needed if the paper parameters are used rather than the re-estimated computer parameters in the CAT administration. This corresponds to a test length that is approximately 29% longer.

An operational testing program might have higher target reliabilities than those defined here, or they might require longer test lengths than the target test lengths defined here. If test lengths were set to be longer than the target test lengths considered here, fewer additional items might be needed to compensate for the loss in reliability due to using paper calibrated parameters with a computer adaptive administration. The predicted test-retest reliabilities are virtually the same for the re-estimated computer parameters and the paper parameters at about a test length of 20 items for Reading Page, and a test length of 24 items for Reading Scroll. Results for Science Reasoning suggest, however, that the minimization of differences in reliability might be dependent on the extent of the parameter differences across modes. For Science Reasoning Scroll, the predicted reliabilities for the paper parameters closely approach (but don't meet) those of the re-estimated computer parameters at about a test length of

33 items. For Science Reasoning Page, even at a test length of 40 items, the predicted reliabilities for the paper parameters are less than those for the re-calibrated computer parameters, although the difference in reliability is relatively small (.013).

Discussion

The evaluation of total scores for Reading showed non-significant mode effects in favor of paper examinees across both the Page and Scroll navigation variations, which translated into small losses in test-retest reliability at some test lengths when using paper calibrated parameters in a CAT administration. The effect of using the paper calibrated parameters appeared similar across the paging and scrolling navigation variations, which suggests that items performed similarly enough across the navigation variations that the CAT administration was not differentially affected. Because the text of the reading passages was very dense and contained minimal white space, it is possible that for some items it was similarly difficult to locate information in the computer presentation across the two methods of navigation.

The evaluation of total scores for Science Reasoning showed larger mode effects in favor of computer examinees, which were significant for the Page condition but not for the Scroll condition. Both navigation variations showed a moderate loss in test-retest reliability when using paper calibrated parameters in a CAT administration, with a more pronounced effect for the Page condition. Unlike the case for Reading, the effect of using paper calibrated parameters appeared somewhat different across the paging and scrolling variations. This suggests that there were large enough parameter differences across the navigation variations to differentially affect the CAT administration. Pommerich (2004) conjectured that computer examinees were better able to focus on the Science Reasoning test than paper examinees, noting that the focus effect could have been aided by the inclusion of figures and tables, which created white space in the text. It is possible that the use of fixed pages as a navigational device might have enabled more of a focus effect on some items than the use of scrolling.

For both methods of navigation, the results from the simulations suggest that using paper calibrated parameters in a fixed length CAT administration could result in test scores that are less reliable than intended. How much less reliable appears dependent on the magnitude of the mode effects and test length. If mode effects are small, test lengths are sufficiently long, and calibration sample sizes are large enough to minimize the effect of estimation error, paper calibrated parameters could probably be used in a CAT administration without incurring too much loss in precision. When mode effects are larger, care should be taken in using paper calibrated parameters in a CAT administration, as greater losses in precision may be incurred that may not be completely recoverable by lengthening the test. Further, given that an often-cited advantage of CAT administration is the ability to meet or exceed the reliability of paper-and-pencil tests with shorter test lengths, lengthening a CAT test to compensate for the use of paper calibrated parameters might be viewed as a less than desirable option.

In closing, it is appropriate to note that the simulation results could have been a little different for both Reading and Science Reasoning, had the not-reached items been treated as not-reached in the calibrations, rather than scored as incorrect. With not-reached items scored as incorrect, items toward the end of the test could have appeared more difficult than they were in reality simply because fewer examinees completed them. If computer examinees could have answered items at the end of the test correctly, but did not reach them because of navigational difficulties, then the simulation results might have appeared less reliable than they really were. This is more likely to be the case in Reading (where completion rates were less for both computer conditions than the paper condition) than in Science Reasoning (where completion rates were higher for both computer conditions than the paper condition).

Endnotes

1. Additional details about the study are available in Pommerich (2004).

2. Table 1 shows sample sizes for the Reading Computer Page and Science Reasoning Computer Page conditions that are smaller than expected, given the random assignment design employed in the data collection. Evaluations of the sample sizes across all of the computer conditions suggest an explanation for this occurrence. In the computer mode, seven different test conditions were administered within each classroom (English Navigation 1, Math, Reading Navigation 1, Science Reasoning Navigation 1, English Navigation 2, Reading Navigation 2, or Science Reasoning Navigation 2). The test administrators were instructed to spiral examinees in order through the seven conditions, and to begin assignments for each new class at the point in the cycle where the previous class had left off. Because the sample counts are largest for the first condition in the sequence (N = 1110), and smallest for the last condition in the sequence (N = 902), with decreasing sample sizes across computer conditions 1–7, it appears that some test administrators may have begun their spiraling assignments for each new class at the first condition rather than continuing where the previous class had left off. The smaller sample sizes for Reading Page and Science Reasoning Page are likely attributable to the fact that they were the sixth and seventh conditions, respectively, in the computer administration cycle.

References

Bennett, R.E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment, 1*(1). Available from http://www.jtla.org.

Choi, S.W., & Tinkler, T. (April, 2002). *Evaluating comparability of paper-and-pencil 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.

Debell, M. & Chapman, C. (2006). *Computer and internet use by students in 2003* (NCES 2006-065). U.S. Department of Education. Washington, D.C.: National Center for Education Statistics.

Hetter, R.D., Segall, D.O., & Bloxom, B.M. (1997). Evaluating item calibration medium in computerized adaptive testing. In W.A. Sands, B.K. Waters, and J.R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (Chapter 16, pp. 161–167). Washington, DC: American Psychological Association.

Hetter, R.D., & Sympson, J.B. (1997). Item exposure control in CAT-ASVAB. In W.A. Sands, B.K. Waters, and J.R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (Chapter 13, pp. 141–144). Washington, DC: American Psychological Association.

Kolen, M.J. (1999-2000). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment, 6*, 73–96.

Paek, P. (2005). Recent trends in comparability studies. PEM Research Report No. 05-05. Pearson Educational Measurement.

Parshall, C.G., Spray, J.A., Kalohn, J.C., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2*(6). Available from http://www.jtla.org.

Sympson, J.B. & Hetter, R.D. (1985). *Controlling item-exposure rates in computerized adaptive testing.* Paper presented at the annual meeting of the Military Testing Association, San Diego, CA.

U.S. Department of Commerce (2002). *A nation online: How Americans are expanding their use of the Internet*. Washington, DC: Author.

Wang, T., & Kolen, M.J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38*, 19–49.

Author Note

The data used in this study was collected while the author was at ACT, Inc. The views expressed are those of the author and not necessarily those of the Department of Defense, the United States Government, or ACT, Inc.

The author thanks Dan Segall for his assistance with the simulation portion of the study.

Correspondence concerning this article should be addressed to:
Mary Pommerich
Defense Manpower Data Center
DoD Center Monterey Bay
400 Gigling Rd.
Seaside, CA 93955-6771
email: mary.pommerich@osd.pentagon.mil

Author Biography

Mary Pommerich is a psychometrician with the ASVAB testing program. She is interested in the pursuit of quality measurement.

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