

■ **The Impact of Item Location Effects on Ability Estimation in CAT: A Simulation Study**

Mei Liu

**Renbang Zhu
Fanmin Guo
Educational Testing Service**

■ **Law School Admission Council
Computerized Testing Report 01-06
September 2006**

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

© 2006 by Law School Admission Council, Inc.

All rights reserved. No part of this report may be reproduced or transmitted in any part or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, 662 Penn Street, Box 40, Newtown, PA 18940-0040.

LSAT and LSAC are registered marks of the Law School Admission Council, Inc.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary	1
Introduction	2
Method	3
Results	7
<i>Impact of Shifted-b Items on Score Estimation</i>	8
<i>Routing Effects and Impact</i>	17
Discussion	29
References.	30
Authors Note	32

Executive Summary

The past decade has seen an increasing number of testing institutions implement large-scale high-stakes computerized adaptive testing (CAT) programs. In these CAT programs, test takers are presented with items selected from an item pool that are tailored to their ability levels while at the same time satisfying content and other psychometric specifications. Each of these collections of items administered to a test taker is really a form or an edition of the computerized adaptive test. Most CAT programs rely on the use of a mathematical model called item response theory (IRT), which makes it possible to compare the proficiency level of test takers even though they have responded to different collections of items.

One of the fundamental assumptions of IRT—local independence—posits that the statistical characteristics of an item are independent of the other items surrounding it in the test. The order of item presentation is one aspect of this assumption. That is, the psychometric characteristics of items (such as item difficulty and discrimination power) will remain invariant regardless of when and where they appear in a test. Meeting this assumption adequately is vital in CAT, since items do not appear in fixed positions as they do in paper-and-pencil tests. Moreover, selecting items dynamically for a test taker during an adaptive test session means that there is no easy way of controlling when and where an item appears. It is therefore imperative that the initial estimates of item characteristics obtained from pretesting continue to hold in an operational CAT environment.

Most testing organizations obtain estimates of item characteristics for their CAT programs by pretesting (or trying out) items in pre-assembled sections either as an external test section or as item groups seeded throughout an operational test section. This means that the items are administered in fixed positions. When these items become operational, their delivery positions can be very different from their pretest delivery positions. The successful implementation of these CAT programs thus rests on the assumption of invariance within reasonable bounds of item characteristics across the pretesting and operational testing environments. If the assumption of item characteristic invariance over item position is not adequately met, the item statistics (such as item difficulty estimates) from pretest data where item position is fixed cannot be generalized to the operational testing environment where item delivery position is unpredictable. Adaptive testing is especially vulnerable to such shifts in item characteristics, because there is no feasible mechanism to neutralize the context effects (e.g., to keep the effects relatively constant for all test takers). Furthermore, the very nature of the adaptive testing process also means that, although all test takers may be affected, they may not be affected in the same way or to the same degree (differential impact).

The purpose of this study was to evaluate the impact of item difficulty shift on ability estimation in CAT as a result of change in item delivery positions from pretest to operational administrations. Three factors were investigated: (1) the direction of the shift (items either became easier or more difficult), (2) the degree of the shift in item difficulty, and (3) the pretest and operational delivery positions of these shifted-difficulty items.

A number of questions were addressed in this study. First, it examined how ability estimation was affected by each of the studied factors as well as the number of shifted-difficulty items simulated test takers encountered in their CAT sessions. In addition, the consistency of the results across ability groups was studied. How the algorithm routed through a CAT pool after items with a shift in difficulty had been introduced was also investigated.

The results indicate that a large shift in item difficulty produced more score impact than a lesser shift in item difficulty. In addition, more impact was observed when the studied items were made easier than when these items were made harder. Furthermore, simulated test takers were likely to experience more impact as the number of shifted-difficulty items they encountered increased. This was particularly evident when these items appeared early in a test. Finally, early appearance of the shifted-difficulty items led to large routing effects, which in turn seemed to produce more scores that differed from their baseline scores by 5 points or more.

Even though these findings demonstrate that multiple items with a shift in item difficulty did not result in impact as severe as that produced by a single flawed item (see Liu & Steffen, 1999), accumulation of smaller effects from these items could still adversely impact some test takers' scores. Moreover, the impact was not uniform across the data conditions investigated in this study. The severity of the impact was influenced by the complex interactions of the ability level of the test takers, the specific shifted-difficulty items they saw, the number of shifted-difficulty items they received, when and where they encountered these items, as well as the other items administered during their testing sessions. Such differential effect, albeit affecting a very small group of test takers, could be problematic if it occurred operationally.

The research presented here relates directly to computerized item-by-item adaptive testing environments. When using more constrained adaptive testing procedures, context effects such as item position can be mitigated or even kept constant. For example, within the multiple-form structure (MFS) methodology that is being investigated at Law School Admission Council (LSAC), intact forms can be calibrated (or tried out) before they are delivered operationally, much as the current paper-and-pencil forms are.

Introduction

The past decade has seen an increasing number of testing institutions implement large-scale high-stakes computerized adaptive testing (CAT) programs. In these CAT programs, test takers are presented with items selected from an item pool that are tailored to their ability levels while at the same time satisfying content and other psychometric specifications. Each of these collections of items administered to a test taker is really a form or an edition of the computerized adaptive test. Most of the CAT programs rely on the use of item response theory (IRT), which makes it possible to calibrate items onto a common scale of difficulty that reflects the underlying, unobserved trait(s) or skill(s). The proficiency levels of test takers taking different collections of items tailored to their ability levels can then be directly placed on the same latent continuum. In other words, the tests drawn from an adaptive item pool are internally equated through the proficiency scale by using the calibrated item parameters so the scores are directly comparable. A well-designed and implemented CAT can provide more reliable and precise measurement of a test taker's ability with fewer items than a traditional paper-and-pencil test. These advantages, however, can be obtained only when the assumptions underlying the model are met reasonably well and the fit between the IRT model and test data is satisfactory.

One of the fundamental assumptions of IRT—local independence—posits that the order of item presentation (one facet of context effects) is irrelevant. That is, item parameters will remain invariant regardless of when and where they appear in a test. Meeting this assumption adequately is vital in CAT, since items do not appear in fixed positions as they do in paper-and-pencil tests. Moreover, selecting items dynamically for a test taker during an adaptive test session means that there is no easy way of controlling when and where an item appears. It is therefore imperative that the initial parameter estimates obtained from pretesting continue to hold in an operational CAT environment. Even though the prevailing view has been that item parameter estimates derived from IRT methods are relatively robust to changes in item position and context, there is some evidence to indicate that parameter estimates are actually sensitive to these context effects (Brennan, 1992; Eignor, 1985; Kingston & Dorans, 1984; Kolen & Harris, 1990; Zwick, 1991).

When test development specialists assemble a new paper-and-pencil test form, they typically follow assembly rules that constrain item positions in test forms to be relatively close to their pretest positions. For example, items that have been pretested in the last few positions cannot be put at the beginning of a new test form. Likewise, items that have been pretested at the beginning of the test cannot appear at the end of a test form. This common practice is followed in order to minimize the effects of changes in context on item parameter estimates between pretest and operational administrations. Even in this kind of controlled environment, it is not uncommon to observe variations in item parameter estimates.

Most testing organizations obtain item parameter estimates for their CAT programs by pretesting in pre-assembled sections either as an external test section or as item blocks embedded in an operational CAT. This means that the items are administered in fixed positions. When these items become operational, their delivery positions can be very different from their pretest delivery positions. The successful implementation of these CAT programs thus rests on the assumption of invariance within reasonable bounds of item parameters across the pretesting and operational testing environments. That is, these CAT programs rely on the initial item parameter estimates to continue to provide appropriate characterizations of item functioning in an operational setting. If the assumption of item parameter invariance over item position is not adequately met, the item parameter estimates calibrated from pretest data where item position is fixed cannot be generalized to the operational testing environment where item delivery position is unpredictable. When this happens, the foundation for adaptive testing becomes shaky. Because items are no longer defined by a common scale of difficulty, scores from different adaptive tests are not comparable anymore, which can lead to incorrect inferences about test takers' proficiencies. In a paper-and-pencil test where every test taker takes the same items in the same order, comparisons among test takers based on IRT scoring are still valid even if the calibrations of items change between pretest and the operational setting due to differences in item location, simply because every test taker encounters the same context. Adaptive testing, on the other hand, is extremely vulnerable to such violations, because there is no feasible mechanism to neutralize the context effects for all test takers (e.g., to keep the effects relatively constant for all test takers). Furthermore, the very nature of the adaptive testing process also means that, although all test takers may be affected, they may not be affected in the same way or to the same degree (differential impact). For example, some test takers may receive more items that are delivered in positions similar to pretest positions than other test takers receive. What happens if some test takers receive more items that exhibit parameter shift due to location effects? Is there a significant differential impact on their scores?

The purpose of this study was to evaluate the impact of item parameter shift on ability estimation in CAT as a result of change in item delivery positions from pretest to operational administrations. A number of questions were addressed in this study. First, it examined how ability estimation was affected by the direction of the parameter shift, the degree of shift, the delivery positions of the items with parameter shift, as well as the number of items with parameter shift that simulated test takers encountered in their CAT sessions. In addition, the consistency of the results across ability groups and how the algorithm routed

through a CAT pool after encountering item(s) with parameter shift were also investigated. This was an exploratory study which examined some what-if scenarios using data derived from simulations in which the true ability level of the simulated test taker was known. It is hoped that the results will provide some food for thought and discussion in the CAT arena.

Method

A retired (formerly operational) CAT pool from a large-scale testing program was used to generate the data. The algorithm used was an adaptation of the weighted deviations model proposed by Stocking & Swanson (1993) and Swanson & Stocking (1993). This algorithm allows the construction of fixed-length adaptive tests by considering the joint contribution of psychometric and content specifications in the item selection process. In addition, the multinomial conditional exposure control method proposed by Stocking and Lewis (1995) was implemented to constrain item use. The data consisted of dichotomous item responses according to a three-parameter (3-PL) logistic IRT model.

Based on the research questions of this study, a three-factor crossed design was used. These factors are shown in Table 1. The first two experimental factors were the direction and the degree of the shift between pretest and operational b-parameter estimates. Twenty items in the CAT pool were manipulated to exhibit b-parameter shift (studied items or shifted-b items). Two levels of shift direction were investigated. All 20 items were manipulated to make them either easier or harder. Four levels of degree of shift were included. A small shift was introduced by adding/subtracting 0.1 from the pretest b-parameter estimates for all studied items. A medium shift was implemented by adding/subtracting 0.15 from the pretest b-parameter estimates for these items, and a large shift was introduced by adding/subtracting 0.25. The fourth level of this factor introduced one-third of the studied items with a small shift, one-third with a medium shift, and one-third with a large shift (combination shift condition).

The third factor in this study was the delivery positions of the studied items when they were administered in pretest and in operational CATs. It is important to note that we did not manipulate the delivery positions of these studied items. Rather, we identified items whose operational delivery positions varied from their pretest positions for a wide range of ability levels from operational data as well as baseline simulation data (item response data simulated without introducing any manipulation). Once items were found to have these characteristics, the shift in b-parameter was introduced wherever these items happened to appear. Three paired (pretest vs. operational) delivery position groups were examined. The first item position group (BA) contained items that were pretested in the middle third of the test (positions 11–20) but were delivered in the first third of an operational CAT (positions 1–10). The second group (AB) included items that were pretested in the first third of the test (positions 1–10) but appeared in the middle third of operational CATs (positions 11–20). The final group (AC) included items that were pretested in the first third of the test (positions 1–10) but appeared in the last third of operational CATs (positions 21–28). Because so few items were pretested in the last third of the test in this CAT pool, we were not able to include a fourth delivery position group (CA) in our study. Note that the first letter of the item delivery group indicates the range of pretest positions and the second letter indicates the range of operational positions.

TABLE 1
A 2 x 4 x 3 study design

Direction of b-Parameter Shift	Degree of b-Parameter Shift	Item Delivery Positions		
		Group	Pretest	Operational
E: Easier	S: Small—0.10	BA:	11–20	1–10
H: Harder	M: Medium—0.15	AB:	1–10	11–20
	L: Large—0.25	AC:	1–10	21–28
	C: 1/3 S, 1/3 M, 1/3 L			

For each of the three delivery position groups, 20 items from the retired CAT pool were selected as studied items. Table 2 displays the pretest IRT parameter estimates for these items. There was some overlap among items delivered in the AB and AC position groups. Common items are shown in bold in the table. Column Six depicts how the fourth level of the degree of shift—the combination shift—was implemented in the study. It shows which items were manipulated to have a large shift, a medium shift, and a small shift.

TABLE 2
Pretest IRT parameters for items selected to be manipulated

Item ¹	a-Parameter	b-Parameter	c-Parameter	Delivery Position	Degree of Shift
1	0.68	-2.57	0.13	BA	L
2	0.93	-2.07	0.08	BA	S
3	0.93	-1.86	0.13	BA	M
4	0.89	-1.04	0.06	BA	M
5	1.04	-0.90	0.17	BA	L
6	0.84	-0.87	0.07	BA	S
7	0.82	-0.86	0.00	BA	M
8	1.12	-0.73	0.16	BA	L
9	0.95	-0.66	0.00	BA	M
10	1.08	-0.08	0.10	BA	L
11	1.22	0.06	0.11	BA	L
12	1.55	0.26	0.17	BA	S
13	1.30	0.60	0.17	BA	S
14	1.34	0.69	0.06	BA	S
15	1.29	0.74	0.19	BA	L
16	1.34	1.12	0.31	BA	L
17	1.09	1.27	0.26	BA	M
18	1.65	1.39	0.11	BA	S
19	1.60	1.67	0.12	BA	S
20	1.25	1.82	0.09	BA	M
1	0.82	-1.72	0.13	AB	M
2	0.75	-1.57	0.00	AB	L
3	0.81	-1.24	0.00	AB	L
4	1.02	-1.20	0.02	AB	M
5	0.88	-1.12	0.06	AB	S
6	0.86	-0.90	0.04	AB	M
7	1.10	-0.75	0.10	AB	S
8	1.05	-0.42	0.05	AB	S
9	1.14	-0.36	0.19	AB	S
10	1.13	-0.02	0.08	AB	L
11	1.38	0.22	0.07	AB	M
12	1.32	0.24	0.07	AB	S
13	1.31	0.28	0.08	AB	L
14	1.44	0.45	0.32	AB	M
15	1.49	0.64	0.18	AB	L
16	0.83	1.10	0.01	AB	M
17	0.91	1.18	0.10	AB	S
18	1.81	1.20	0.20	AB	M
19	1.55	1.22	0.12	AB	L
20	1.19	1.38	0.23	AB	S
1	0.67	-1.18	0.18	AC	S
2	0.81	-1.01	0.10	AC	S
3	0.86	-0.93	0.05	AC	M
4	1.02	-1.20	0.02	AC	M
5	0.61	-0.84	0.24	AC	L
6	0.86	-0.90	0.04	AC	L
7	0.61	-0.34	0.17	AC	M
8	1.05	-0.42	0.05	AC	L
9	1.14	-0.36	0.19	AC	M
10	1.08	-0.21	0.30	AC	M
11	0.83	0.14	0.05	AC	L
12	1.26	0.14	0.01	AC	S
13	1.34	0.21	0.20	AC	L
14	0.81	0.35	0.33	AC	S
15	1.49	0.64	0.18	AC	M
16	0.83	1.10	0.01	AC	M
17	1.20	0.43	0.12	AC	L
18	1.28	0.55	0.29	AC	S
19	1.13	0.59	0.07	AC	S
20	1.19	1.38	0.23	AC	L

¹Items that appeared in both the AB and AC position groups are indicated in bold.

The data in this study were generated by mimicking what would happen in an operational setting when some item parameters from pretest fail to hold in later test administrations. Items were considered to have two sets of operating parameters: (1) the actual parameters from pretest that were used during the item selection and scoring process, and (2) the shifted parameters that were used to generate item responses. For items that were not manipulated (non-studied items), the two sets of item parameters were identical. The parameter shift was introduced by adding or subtracting a constant (0.10, 0.15, and 0.25) from the pretest b-parameter estimates. Figure 1 displays three graphs of item characteristic curves (ICC) of a studied item. Each graph shows three ICCs. The solid-line curve (middle curve) is based on pretest item parameter estimates; the dash-line curve (upper curve) is based on the item parameter estimates with an easy shift; the dotted-line curve (lower curve) is based on the item parameter estimates with a hard shift. For the baseline data, the middle curve was used to generate item responses for the item selection and scoring process. For the shifted-b data, the middle curve was used for the item selection and scoring process. The item responses to the non-studied items were generated using the middle curve, whereas the item responses to the studied items were generated by using either the upper curve (easy shift) or the lower curve (hard shift). All data conditions within an item delivery position group employed the same studied items.

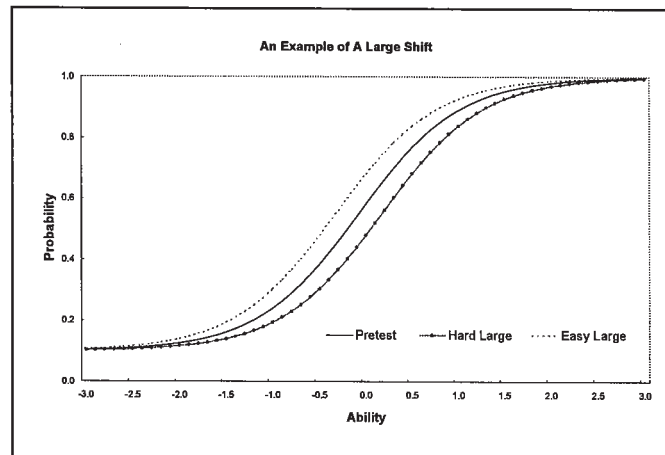
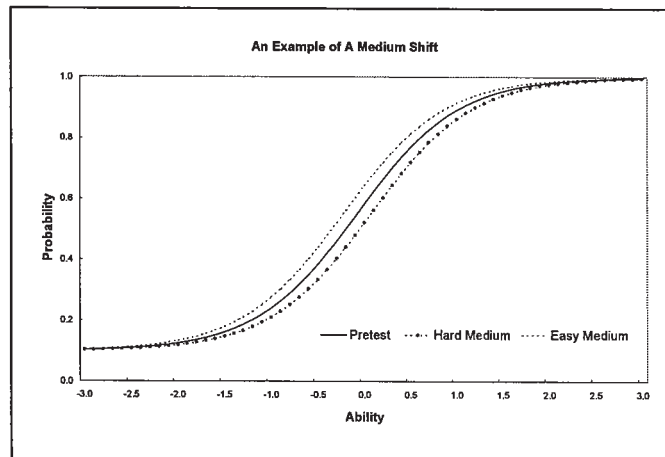
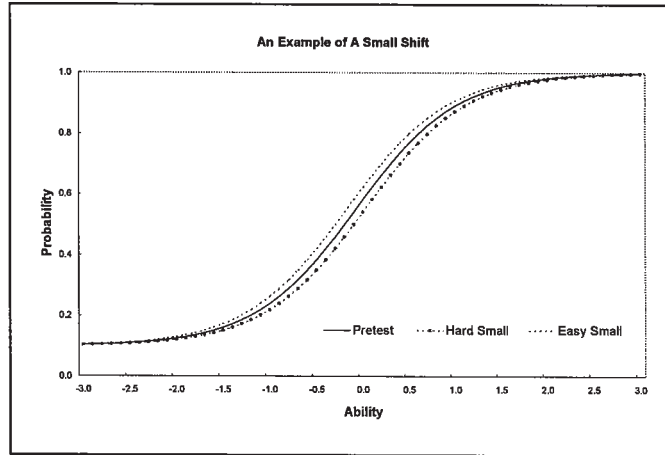


FIGURE 1. Degree of shift introduced to studied items

Two kinds of simulations were performed. A baseline simulation was run in which no item was manipulated. In other words, the generated response data were consistent with the item parameters used for item selection and scoring during a test administration. This baseline data would be used as a criterion to evaluate the data derived from the second kind of simulations—the shifted-b simulations—in which the b-parameters of the studied items were manipulated to generate the item response data. As a result, the shifted-b data were not entirely consistent with the item parameters used for item selection and scoring during a CAT administration. The three factor ($2 \times 4 \times 3$) design employed in this study resulted in a total of 24 shifted-b simulation runs.

The outcome score for the data in this study was the predicted number-right score on a 60-item reference test. Five thousand cases were generated at each reported score level in order to have sufficient data for analyses performed by ability levels. The CAT consisted of 28 items. To make the data more comparable, the same random seed was used for the baseline and the shifted-b data conditions.

Before we get to the result section, it is necessary to keep in mind the difference between a study employing a completely experimental design and one employing a partially experimental design. A completely experimental study is better able to systematically manipulate the data and isolate various effects. Sometimes, however, the systematic data manipulation might change the dynamics of an environment and make the results less generalizable to real world situations. A partially experimental design may only be able to manipulate the data to a certain degree and is often used to study phenomena occurring in less controlled settings. The conclusions from such studies, while not as strong as those based on completely experimental designs, tend to describe the real world better. These two types of designs complement each other, and using both to study a problem whenever feasible can be very beneficial.

To investigate the effects of parameter shift on IRT ability estimation due to change in item delivery positions between pretest and operational CAT administrations, one can approach it by directly manipulating the parameter shift as well as the item delivery positions. That is, the researchers decide in advance where in a CAT the studied items are to appear. While such a study has better control in item delivery positions of the studied items (they can only appear in positions pre-determined for them), it interferes with how the underlying CAT algorithm operates in real testing sessions and changes the dynamics of the item pool.

This research employed a partially experimental design. While a parameter shift was introduced to the studied items, it was introduced to these items in their naturally occurring delivery positions. By not overriding the underlying CAT algorithm, we were able to observe more realistically how it reacted to the circumstances when parameter shift occurred. Employing such a partially experimental design for this study has its limitations, though. The biggest one is that we could not use the same set of studied items across the three delivery position groups, which makes comparisons of results across item delivery positions more difficult. On the other hand, this is exactly the kind of situation an operational CAT testing program has to reckon with on a daily basis. The next difficulty is related to multiple delivery positions. Since it is not uncommon for an item to appear in multiple delivery positions, some simulated test takers received some of the studied items in delivery positions other than those examined in this study (target positions). For example, some of the studied items in the AC delivery positions might also appear in the beginning two-thirds of a CAT. Likewise, some studied items in the BA delivery positions could show up in the last one-third of a CAT. To reduce the potential confounding effect due to multiple delivery locations, results were analyzed by only including those simulated test takers who received the studied item(s) in the three target position groups. As a result, a sufficient number of response cases had to be generated at each ability level to compensate for the data loss.

Results

To ensure more meaningful comparisons between the baseline and the shifted-b data conditions, only simulated test takers who received studied item(s) in the three target delivery position groups in both baseline and shifted-b conditions were included in the analyses reported in this section.

The results presented in this section will also be compared to those of a study by Liu and Steffen (1999). In that study, the impact of a single flawed item introduced in 10 delivery positions of a CAT on item routing and ability estimation was investigated. Three of the delivery positions were located in the first third of the test, three were located in the middle third of the test, and four were located in the last third of the test. The flawed item was introduced in each data condition by reverse keying it. Items with shifted-b parameters can also be considered as flawed items in the sense that their parameter estimates no longer adequately describe the data. The comparisons of these findings should provide helpful perspectives, since this study introduced multiple items that were somewhat flawed (easier or harder than they actually were), whereas the 1999 study manipulated a single item that was seriously flawed. In that study, the probability of getting an item correct was completely reversed.

Impact of Shifted-b Items on Score Estimation

Impact in this study is defined as the difference between the estimated true score under the baseline condition where no parameter shift was introduced and the estimated true score under the shifted-b condition. A rounded score difference of zero indicates no impact. The true scores are on a scale of 0 to 60. The conditional standard errors of measurement range from 3 to 4.

Table 3 reports the means and standard deviations of impact for the 24 data conditions. The first and second columns indicate the direction and the degree of shift introduced to the b-parameters. The direction of impact is as expected. CAT forms containing items that were manipulated to be easier were on average easier than their corresponding baseline forms as shown by the negative means, indicating that the shifted-b CATs produced higher means than the baseline forms. CAT forms with items that were shifted to be harder were more difficult than the baseline CATs on average. The biggest impact is observed in CAT forms that contained items with a large shift. Overall, the CAT forms with a small shift were affected the least.

TABLE 3

Means and standard deviations of impact¹—Simulated test takers who received studied item(s) in target positions of both baseline and shifted-b data conditions

Shift Direction	Shift Degree	Item Delivery Positions								
		BA ²			AB ³			AC ⁴		
		Mean	SD	N	Mean	SD	N	Mean	SD	N
Easier	Lg, Med, Sm	-0.165	0.700	69,991	-0.143	0.557	10,146	-0.101	0.391	40,307
Easier	Large	-0.274	0.873	71,284	-0.222	0.688	10,155	-0.155	0.487	40,307
Easier	Medium	-0.181	0.723	70,267	-0.140	0.553	10,177	-0.099	0.390	40,307
Easier	Small	-0.133	0.632	69,722	-0.097	0.470	10,175	-0.071	0.332	40,307
Harder	Lg, Med, Sm	0.057	0.590	67,401	0.127	0.546	10,272	0.074	0.391	40,307
Harder	Large	0.138	0.728	66,026	0.206	0.678	10,266	0.126	0.482	40,307
Harder	Medium	0.066	0.606	67,130	0.116	0.530	10,240	0.068	0.382	40,307
Harder	Small	0.031	0.532	67,752	0.079	0.454	10,243	0.041	0.323	40,307

¹Impact = Estimated True Score From Baseline Data Condition - Estimated True Score From Shifted-b Data Condition.

²Items pretested in positions 11–20 but appeared in positions of 1–10 in operational CATs.

³Items pretested in positions 1–10 but appeared in positions of 11–20 in operational CATs.

⁴Items pretested in positions 1–10 but appeared in positions of 21–28 in operational CATs.

For the easy shift data conditions, the largest impact occurred for the BA delivery position group in which the studied items were pretested in positions 11 through 20 then administered in positions 1 through 10 in the operational CATs. CAT forms that contained the studied items in the AC delivery position group (pretest positions 1–10, operational positions 21–28) were the least impacted. CAT forms with the studied items in AB delivery position group (pretest positions of 1–10, operational positions 11–20) fell in between. For the hard shift data conditions, the most impacted were CAT forms that contained the studied items in the AB delivery position group. The forms in the AC and BA position groups experienced similar impact. Impact scores of the BA position groups were more variable than those of the other two groups.

The difference in means between the easy shift and hard shift data conditions for the three delivery position groups is a little surprising. This difference is much bigger for the BA position group than for the other two delivery position groups. For example, for the BA delivery position group under the easy shift conditions, the CAT forms on average were about 0.13 to 0.27 of a true score point easier than their baseline scores. When the items were made harder, the CAT forms on average were only about 0.03 to 0.14 of a true score point harder than the baseline CAT forms. For the AB delivery position group, means ranged from -0.10 to -0.22 for the easy shift conditions, and 0.08 to 0.21 for the hard shift conditions.

Tables 4 and 5 show the means and standard deviations of impact by true score level for the large and small shift data conditions. The first column displays the direction and degree of the shift. The second column contains the true score that is grouped into 10 intervals. As expected, impact is related to ability level, direction of shift, as well as delivery positions. When items were manipulated to be easier, the highest ability group (true score 55 or higher) was the least impacted. The lowest ability group (true score between 10 and 14) was affected the most (except for AC position group), followed by the middle six ability groups (true score ranges from 20 to 49). Impact also tends to be less variable for the highest ability group. The middle six ability groups (true score ranges from 20 to 49) tended to experience more impact in the BA and AB delivery positions than those in the AC delivery positions. When items were manipulated to become harder, the pattern was somewhat different. The lowest ability group was still the most affected (except for the AC group). But the highest ability group was no longer the least affected across the three delivery positions. In fact the highest ability group was the second most affected in the BA position. The middle six ability groups were affected more in the AB and in the AC delivery positions than in the BA delivery positions (except for true score levels of 20–24 and 45–49).

What seems counterintuitive is the direction of impact for the lowest ability group when items were made more difficult. For this group, the CAT forms became easier on average instead of becoming harder than their baseline forms. This is likely due to the fact that the probability of answering the studied items correctly was decreased as compared to the baseline condition, because these items were manipulated to become harder, which resulted in easier forms being administered to this group of simulated test takers. Encountering the studied items early as in the AB and BA delivery positions resulted in larger impact because there were more slots for these simulated test takers to see easier items than later in their CATs as demonstrated by the results for the AC delivery position group.

TABLE 4

Means and standard deviations of impact¹ by true score level for selected data conditions—Simulated test takers who received studied item(s) in target positions of both baseline and shifted-b data conditions

Direction and Degree of b Shift	True Score Range	Item Delivery Positions								
		BA			AB			AC		
		Mean	SD	N	Mean	SD	N	Mean	SD	N
Easier Large	10–14	–0.812	1.078	5,084	–0.743	1.139	144	–0.146	0.402	5,914
	15–19	–0.183	0.622	3,230	–0.174	0.528	298	–0.090	0.332	3,542
	20–24	–0.265	0.912	2,253	–0.190	0.766	564	–0.193	0.556	3,676
	25–29	–0.262	1.042	3,087	–0.151	0.572	736	–0.185	0.572	3,655
	30–34	–0.204	0.885	5,708	–0.253	0.788	1,177	–0.220	0.625	4,368
	35–39	–0.324	1.101	6,666	–0.301	0.835	1,646	–0.272	0.666	4,651
	40–44	–0.259	0.950	7,515	–0.238	0.681	1,564	–0.190	0.529	3,754
	45–49	–0.222	0.860	10,174	–0.236	0.677	1,405	–0.097	0.352	4,200
	50–54	–0.303	0.835	15,115	–0.170	0.543	1,914	–0.059	0.247	4,569
	55 & Up	–0.105	0.462	12,452	–0.075	0.330	707	–0.033	0.205	1,978
Easier Small	10–14	–0.692	1.025	5,086	–0.559	1.025	143	–0.118	0.371	5,914
	15–19	–0.077	0.410	3,303	–0.055	0.293	293	–0.032	0.212	3,542
	20–24	–0.121	0.620	2,248	–0.102	0.610	566	–0.079	0.365	3,676
	25–29	–0.089	0.646	2,984	–0.056	0.372	728	–0.061	0.340	3,655
	30–34	–0.067	0.532	5,570	0.072	0.423	1,162	–0.072	0.367	4,368
	35–39	–0.136	0.742	6,430	–0.134	0.563	1,643	–0.130	0.466	4,651
	40–44	–0.111	0.622	7,302	–0.096	0.436	1,581	–0.085	0.365	3,754
	45–49	–0.085	0.530	10,022	–0.106	0.457	1,412	–0.039	0.226	4,200
	50–54	–0.140	0.591	14,588	–0.093	0.424	1,936	–0.026	0.166	4,569
	55 & Up	0.000	0.414	12,189	–0.011	0.265	711	–0.005	0.188	1,978

¹Impact = Estimated True Score From Baseline Data Condition – Estimated True Score From Shifted-b Data Condition.

TABLE 5

Means and standard deviations of impact¹ by true score level for selected data conditions—Simulated test takers who received studied item(s) in target positions of both baseline and shifted-b data conditions

Direction and Degree of b Shift	True Score Range	Item Delivery Positions								
		BA			AB			AC		
		Mean	SD	N	Mean	SD	N	Mean	SD	N
Harder Large	10–14	–0.415	0.857	5,073	–0.295	0.807	146	–0.060	0.256	5,914
	15–19	0.204	0.622	3,446	0.234	0.689	286	0.074	0.336	3,542
	20–24	0.182	0.671	2,292	0.191	0.711	554	0.138	0.514	3,676
	25–29	0.123	0.702	2,755	0.190	0.759	716	0.198	0.600	3,655
	30–34	0.142	0.757	5,296	0.226	0.755	1,165	0.232	0.624	4,368
	35–39	0.169	0.790	6,070	0.220	0.729	1,681	0.228	0.609	4,651
	40–44	0.152	0.761	6,740	0.257	0.710	1,620	0.212	0.562	3,754
	45–49	0.149	0.725	9,658	0.244	0.696	1,426	0.142	0.437	4,200
	50–54	0.182	0.665	13,380	0.177	0.533	1,945	0.066	0.269	4,569
	55 & Up	0.273	0.592	11,316	0.153	0.420	727	0.083	0.291	1,978
Harder Small	10–14	–0.530	0.935	5,086	–0.444	0.944	144	–0.078	0.293	5,914
	15–19	0.079	0.426	3,390	0.069	0.346	291	0.034	0.265	3,542
	20–24	0.073	0.426	2,291	0.106	0.591	559	0.048	0.321	3,676
	25–29	0.054	0.469	2,855	0.077	0.482	718	0.083	0.383	3,655
	30–34	0.073	0.538	5,392	0.120	0.561	1,167	0.113	0.448	4,368
	35–39	0.054	0.487	6,266	0.077	0.426	1,667	0.071	0.346	4,651
	40–44	0.054	0.485	6,942	0.100	0.451	1,607	0.083	0.360	3,754
	45–49	0.060	0.472	9,852	0.098	0.450	1,421	0.051	0.269	4,200
	50–54	0.059	0.392	13,902	0.062	0.326	1,949	0.020	0.147	4,569
	55 & Up	0.141	0.437	11,776	0.064	0.272	720	0.040	0.213	1,978

¹Impact = Estimated True Score From Baseline Data Condition - Estimated True Score From Shifted-b Data Condition.

A similar pattern was also observed for the highest ability group when items were manipulated to be easier with small shift. The CAT forms in the shifted-b conditions for this group were on average as difficult as the baseline forms. Given the type of item pool used in this study, this group of simulated test takers was expected to answer correctly almost every item presented. Making some of the items they encountered slightly easier would not have changed the difficulty level of subsequent items they received, which resulted in CAT forms that were of comparable difficulty level as the baseline forms.

Table 6 presents the frequency distributions of impact for the three delivery position groups. Impact is grouped into several categories: no impact, small impact of 1 to 2 true score points, moderate impact of 3 and 4 score points, and large impact of 5 or more score points. The first column shows the delivery position group. The second and third columns display the direction and degree of shift. The next four columns present the percent of simulated test takers who experienced impact.

TABLE 6

Distributions of absolute impact¹—Simulated test takers who received studied item(s) in target positions for both baseline and shifted-b conditions

Item	Direction of b Shift	Degree of b Shift	Absolute Difference in Estimated True Scores				N
			0 %	1-2 %	3-4 %	5 & Up %	
BA	Easier	L,M,S	88.5	9.3	1.8	0.4	69,991
		Large	82.7	13.6	3.1	0.7	71,284
		Medium	87.6	10.0	2.0	0.4	70,267
		Small	90.2	8.1	1.4	0.3	69,722
	Harder	L,M,S	90.5	8.1	1.1	0.2	67,401
		Large	86.2	11.6	1.9	0.3	66,026
		Medium	90.1	8.4	1.3	0.2	67,130
		Small	92.1	6.8	0.9	0.2	67,752
AB	Easier	L,M,S	91.3	7.4	1.3	0.1	10,146
		Large	86.9	10.9	2.0	0.2	10,155
		Medium	91.5	7.2	1.2	0.1	10,177
		Small	94.0	5.1	0.9	0.1	10,175
	Harder	L,M,S	91.5	7.4	1.0	0.1	10,272
		Large	86.7	11.3	1.8	0.2	10,266
		Medium	92.1	6.8	0.9	0.1	10,240
		Small	94.5	4.7	0.7	0.1	10,243
AC	Easier	L,M,S	92.2	7.4	0.3	0.0	40,307
		Large	88.6	10.7	0.6	0.0	40,307
		Medium	92.4	7.2	0.3	0.0	40,307
		Small	94.5	5.3	0.2	0.0	40,307
	Harder	L,M,S	92.8	6.9	0.3	0.0	40,307
		Large	89.7	9.7	0.6	0.0	40,307
		Medium	93.4	6.2	0.3	0.0	40,307
		Small	95.1	4.7	0.2	0.0	40,307

¹Absolute Difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

Impact is slightly larger for the BA and AB positions than for the AC delivery position. Between 83% and 95% of the simulated test takers did not experience any impact across all three delivery position groups. Less than 1 % of the simulated test takers experienced a score change of three to four true score points when they encountered shifted-b items in the AC delivery position group. For the BA position group, however, about 1% to 4% of the simulated test takers experienced moderate to large score impact. For the AB position group, about 1 % to 2% of the simulated test takers saw moderate to large score changes. Regardless of the direction of the shift, a large shift resulted in the most impact, followed by the combination, the medium (except for the BA group) and the small shift data conditions. Easy shift conditions produced slightly more impact than hard shift conditions.

The results reported so far are consistent with what has been reported in the flawed-item study by Liu and Steffen (1999). In that study, it was observed that the impact was larger when a flawed item appeared early in a CAT. Impact in this study was not as severe as in the flawed-item cases. Fewer simulated test takers experienced large score changes of five points or more.

To see where on the true score scale these score differences occurred, distributions of absolute impact were examined by true score level. Tables 7 through 12 present these distributions for both shift directions of the three delivery position groups. The first column shows the true score levels, and the next four columns display the percent of simulated test takers who experienced score change from baseline to the shifted-b data conditions. Tables 7, 8, and 9 present the results for the large shift data conditions, and Tables 10, 11, and 12 show the results for the small shift data conditions.

TABLE 7

Distribution of absolute impact¹ by true score level for selected data conditions—Items pretested in positions 11–20 but appeared in positions 1–10 in operational CATs (BA)—Simulated test takers who received studied item(s) in BA positions for both baseline and shifted-b conditions, items became easier

True Score Range	Absolute Difference in Estimated True Scores				N
	0 %	1–2 %	3–4 %	5 & Up %	
	Items Become Easier (E) with Large Shift (L)				
10–14	50.4	42.6	6.5	0.6	5,084
15–19	86.7	12	1.2	0.2	3,230
20–24	87.1	8.7	3.1	1.1	2,253
25–29	88.1	6.5	3.6	1.7	3,087
30–34	89.2	6.9	2.9	0.9	5,708
35–39	83.2	10.3	5.0	1.5	6,666
40–44	85.6	9.7	3.7	1.0	7,515
45–49	86.2	9.9	3.3	0.6	10,174
50–54	79.2	17.3	3.1	0.4	1,5115
55 & Up	89.2	10.3	0.5	0.0	12,452
	Items Become Easier (E) with Small Shift (S)				
10–14	56.3	38.1	5.1	0.5	5,086
15–19	93.9	5.7	0.4	0.0	3,303
20–24	93.7	4.6	1.1	0.5	2,248
25–29	96.0	2.0	1.4	0.6	2,984
30–34	96.4	2.4	1.0	0.3	5,570
35–39	92.5	4.8	2.1	0.7	6,430
40–44	93.7	4.4	1.4	0.5	7,302
45–49	94.7	3.9	1.3	0.2	10,022
50–54	90.0	8.4	1.4	0.2	14,588
55 & Up	91.6	7.9	0.4	0.0	12,189

¹Absolute Difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

TABLE 8

Distribution of absolute impact¹ by true score level for selected data conditions—Items pretested in positions 11–20 but appeared in positions 1–10 in operational CATs (BA)—Simulated test takers who received studied item(s) in BA positions for both baseline and shifted-b conditions, items became harder

True Score Range	Absolute Difference in Estimated True Scores				N
	0 %	1–2 %	3–4 %	5 & Up %	
	Items Become Harder (H) with Large Shift (L)				
10–14	72.3	24.9	2.6	0.2	5,073
15–19	86.1	12.4	1.3	0.1	3,446
20–24	88.9	8.7	2.1	0.2	2,292
25–29	92.2	5.3	1.9	0.5	2,755
30–34	91.9	5.4	2.1	0.6	5,296
35–39	90.9	5.6	2.8	0.6	6,070
40–44	90.6	6.4	2.4	0.6	6,740
45–49	90.1	7.2	2.1	0.5	9,658
50–54	87.0	10.9	1.8	0.2	13,380
55 & Up	78.0	21.1	0.9	0.0	11,316
	Items Become Harder (H) with Small Shift (S)				
10–14	65.3	31.0	3.4	0.3	5,086
15–19	92.5	7.0	0.5	0.0	3,390
20–24	95.1	4.1	0.7	0.1	2,291
25–29	96.4	2.4	0.9	0.2	2,855
30–34	96.1	2.5	1.0	0.3	5,392
35–39	96.6	2.2	1.0	0.2	6,266
40–44	96.2	2.6	1.0	0.2	6,942
45–49	95.7	3.2	1.0	0.2	9,852
50–54	95.5	3.8	0.6	0.1	13,902
55 & Up	88.1	11.6	0.4	0.0	11,776

¹Absolute Difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

TABLE 9

Distribution of absolute impact¹ by true score level for selected data conditions—Items pretested in positions 1–10 but appeared in positions 11–20 in operational CATs (AB)—Simulated test takers who received studied item(s) in AB positions for both baseline and shifted-b conditions, items became easier

True Score Range	Absolute Difference in Estimated True Scores				N
	0 %	1–2 %	3–4 %	5 & Up %	
	Items Become Easier (E) with Large Shift (L)				
10–14	64.6	25.7	9.0	0.7	144
15–19	88.6	10.7	0.7	0.0	298
20–24	88.8	8.7	2.0	0.5	564
25–29	90.6	7.9	1.5	0.0	736
30–34	88.0	8.7	2.8	0.5	1,177
35–39	84.1	12.5	2.8	0.6	1,646
40–44	85.5	12.2	2.2	0.0	1,564
45–49	85.3	12.7	2.0	0.1	1,405
50–54	88.0	10.8	1.0	0.1	1,914
55 & Up	92.9	6.8	0.3	0.0	707
	Items Become Easier (E) with Small Shift (S)				
10–14	73.4	18.9	7.7	0.0	143
15–19	95.9	3.8	0.3	0.0	293
20–24	93.3	5.5	0.9	0.4	566
25–29	96.6	2.7	0.7	0.0	728
30–34	96.6	2.4	1.0	0.0	1,162
35–39	92.6	6.0	1.2	0.1	1,643
40–44	94.1	5.2	0.8	0.0	1,581
45–49	93.6	5.6	0.8	0.0	1,412
50–54	93.7	5.6	0.6	0.1	1,936
55 & Up	95.4	4.5	0.1	0.0	711

¹Absolute Difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

TABLE 10

Distribution of absolute impact¹ by true score level for selected data conditions—Items pretested in positions 1–10 but appeared in positions 11–20 in operational CATs (AB)—Simulated test takers who received studied item(s) in AB positions for both baseline and shifted-b conditions, items became harder

True Score Range	Absolute Difference in Estimated True Scores				N
	0 %	1–2 %	3–4 %	5 & Up %	
	Items Become Harder (H) with Large Shift (L)				
10–14	87.0	8.9	4.1	0.0	146
15–19	87.4	10.5	1.7	0.3	286
20–24	89.5	9.4	0.9	0.2	554
25–29	90.1	7.7	1.7	0.6	716
30–34	86.7	10.6	2.2	0.5	1,165
35–39	87.7	9.8	2.2	0.3	1,681
40–44	84.4	13.2	2.2	0.1	1,620
45–49	85.8	11.2	3.0	0.0	1,426
50–54	86.1	13.1	0.7	0.1	1,945
55 & Up	86.9	12.9	0.1	0.0	727
	Items Become Harder (H) with Small Shift (S)				
10–14	79.2	15.3	5.6	0.0	144
15–19	95.5	4.1	0.3	0.0	291
20–24	95.0	4.5	0.4	0.2	559
25–29	95.5	3.8	0.4	0.3	718
30–34	93.7	4.8	1.2	0.3	1,167
35–39	95.8	3.6	0.5	0.1	1,667
40–44	94.3	4.9	0.8	0.0	1,607
45–49	94.1	4.9	1.1	0.0	1,421
50–54	95.1	4.6	0.3	0.1	1,949
55 & Up	94.2	5.7	0.1	0.0	720

¹Absolute Difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

TABLE 11

Distribution of absolute impact¹ by true score level for selected data conditions—Items pretested in positions 1–10 but appeared in positions 21–28 in operational CATs (AC)—Simulated test takers who received studied item(s) in AC positions for both baseline and shifted-b conditions, items became easier

True Score Range	Absolute Difference in Estimated True Scores				N
	0 %	1–2 %	3–4 %	5 & Up %	
	Items Become Easier (E) with Large Shift (L)				
10–14	86.7	13.1	0.2	0.0	5,914
15–19	92.2	7.7	0.1	0.0	3,542
20–24	86.7	12.3	1.0	0.1	3,676
25–29	88.0	10.7	1.1	0.1	3,655
30–34	86.2	12.3	1.4	0.0	4,368
35–39	82.8	15.8	1.3	0.1	4,651
40–44	86.5	13.0	0.5	0.0	3,754
45–49	92.0	7.9	0.1	0.0	4,200
50–54	94.2	5.8	0.0	0.0	4,569
55 & Up	96.0	4.0	0.0	0.0	1,978
	Items Become Easier (E) with Small Shift (S)				
10–14	89.3	10.4	0.2	0.0	5,914
15–19	97.2	2.8	0.0	0.0	3,542
20–24	94.5	5.1	0.4	0.0	3,676
25–29	96.0	3.6	0.4	0.1	3,655
30–34	95.3	4.3	0.4	0.0	4,368
35–39	91.5	7.9	0.6	0.0	4,651
40–44	94.0	5.9	0.2	0.0	3,754
45–49	96.7	3.2	0.1	0.0	4,200
50–54	97.4	2.6	0.0	0.0	4,569
55 & Up	96.9	3.1	0.0	0.0	1,978

¹Absolute Difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

TABLE 12

Distribution of absolute impact¹ by true score level for selected data conditions—Items pretested in positions 1–10 but appeared in positions 21–28 in operational CATs (AC)—Simulated test takers who received studied item(s) in AC positions for both baseline and shifted-b conditions, items became harder

True Score Range	Absolute Difference in Estimated True Scores				N
	0 %	1–2 %	3–4 %	5 & Up %	
	Items Become Harder (H) with Large Shift (L)				
10–14	94.4	5.6	0.1	0.0	5,914
15–19	93.9	6.0	0.1	0.1	3,542
20–24	90.7	8.5	0.7	0.1	3,676
25–29	87.4	11.1	1.4	0.1	3,655
30–34	85.1	13.5	1.3	0.1	4,368
35–39	85.3	13.5	1.2	0.0	4,651
40–44	85.2	14.1	0.7	0.0	3,754
45–49	88.9	10.9	0.2	0.0	4,200
50–54	93.8	6.2	0.0	0.0	4,569
55 & Up	92.1	7.9	0.0	0.0	1,978
	Items Become Harder (H) with Small Shift (S)				
10–14	92.7	7.2	0.1	0.0	5,914
15–19	97.0	2.9	0.1	0.1	3,542
20–24	96.8	3.0	0.2	0.1	3,676
25–29	94.5	5.0	0.5	0.0	3,655
30–34	92.7	6.8	0.5	0.1	4,368
35–39	95.1	4.6	0.3	0.0	4,651
40–44	94.1	5.7	0.2	0.0	3,754
45–49	95.9	4.0	0.1	0.0	4,200
50–54	98.1	1.9	0.0	0.0	4,569
55 & Up	96.2	3.8	0.0	0.0	1,978

¹Absolute Difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

Clearly a large shift produced more impact except for the lowest ability group in the hard shift data conditions. In general, the highest ability group was the least impacted in terms of moderate to large impact, and the lowest ability group was the most affected, especially with regard to small degrees of impact. No simulated test takers in the highest ability group experienced large impact regardless of the direction of the shift, the degree of the shift, or where the shifted-b items appeared in the test. More simulated test takers from lower to middle ability groups were impacted when items were made easier. Consider the BA delivery positions as an example (Tables 7 and 8). When items were made easier with a large shift, 50% of the simulated test takers from the lowest ability group did not experience any impact, and 7% of the simulated test takers in this ability group received scores that were different from their baseline scores by three or more points (moderate impact). Between 4% and 7% of the middle six ability groups also experienced moderate to large impact. When the shifted-b items were manipulated to be more difficult, however, only 3% of the lowest ability group observed a score change of three or more points, and 2% to 3% of the middle six ability groups saw similar score change. Of the three delivery position groups, BA delivery positions produced the largest impact, whereas the AC delivery positions had the least impact.

One potential concern from these results is that when a large shift was introduced earlier in CATs (e.g., positions BA), about two to five percent of the scores differ from their baseline scores by three or more true score points for the higher ability groups. (True scores range between 40 to 54.) The conditional standard error of measurement is between three and four score points, with four points being the highest. So some of these score differences are real. The important question is whether such score change would lead to different admission/selection decisions by some test takers if similar parameter shift had occurred in operational settings.

Additional analyses were performed to examine how the number of shifted-b items simulated test takers received impacted their scores. Tables 13 through Table 15 report the distributions of absolute impact by the number of shifted-b items simulated test takers received in their CAT forms in the three groups of item delivery positions. Column 1 indicates the direction and the degree of the shift, and column 2 shows the number of shifted-b items each simulated test taker received in the test. Columns 3 through 6 show the percentages of simulated test takers who experienced various degrees of impact on their scores across the eight data combinations. Following the third column in all three tables, one can see that, as the number of shifted-b items increased, the percentages of simulated test takers experiencing no impact decreased. Moreover, the percent of simulated test takers experiencing moderate to large score change increased as the number of shifted-b items delivered to these simulated test takers increased. The sole exception is for the AB condition (Table 14) for items that became slightly harder. Here, there was no large impact for the 12 simulees. However, since very few simulated test takers received five or more shifted-b items at the AB delivery positions, the percentages for this category may not be very stable.

TABLE 13

Distribution of absolute impact¹ by number of items with shifted-b parameters—Items pretested in positions 11–20 but appeared in positions 1–10 of operational CATs (BA)—Simulated test takers who received studied item(s) in BA positions for both baseline and shifted-b conditions

Direction and Degree of b Shift	Number of Items With b Shift	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Easier	1–2	92.1	6.5	1.2	0.2	41,913
Lg, Med, Sm	3–4	83.1	13.6	2.7	0.5	27,547
	5 & Up	80.6	13.0	4.9	1.5	531
Easier	1–2	87.7	9.7	2.1	0.4	42,133
Large	3–4	75.5	19.1	4.5	0.9	28,612
	5 & Up	71.8	19.1	6.3	2.8	539
Easier	1–2	91.5	6.9	1.3	0.3	41,931
Medium	3–4	81.9	14.6	3.0	0.5	27,803
	5 & Up	80.5	12.9	5.1	1.5	533
Easier	1–2	93.5	5.5	0.9	0.2	41,776
Small	3–4	85.3	12.0	2.2	0.4	27,416
	5 & Up	84.7	9.8	4.3	1.1	530
Harder	1–2	93.0	6.0	0.8	0.2	41,209
Lg, Med, Sm	3–4	86.8	11.3	1.7	0.3	25,696
	5 & Up	84.3	12.1	2.6	1.0	496
Harder	1–2	89.2	9.1	1.5	0.3	40,841
Large	3–4	81.3	15.7	2.6	0.4	24,690
	5 & Up	77.2	17.0	4.2	1.6	495
Harder	1–2	92.7	6.2	0.9	0.2	41,126
Medium	3–4	86.1	11.8	1.8	0.3	25,498
	5 & Up	83.6	12.1	3.4	1.0	506
Harder	1–2	94.4	4.9	0.6	0.1	41,289
Small	3–4	88.5	9.9	1.4	0.2	25,961
	5 & Up	87.5	9.2	2.4	1.0	502

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

TABLE 14

Distribution of absolute impact¹ by number of items with shifted-b parameters—Items pretested in positions 1–10 but appeared in positions 11–20 of operational CATs (AB)—Simulated test takers who received studied item(s) in AB positions for both baseline and shifted-b conditions

Direction and Degree of b Shift	Number of Items With b Shift	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Easier	1–2	92.9	6.0	1.0	0.1	7,890
Lg, Med, Sm	3–4	85.5	12.2	2.1	0.1	2,246
	5 & Up	70.0	20.0	10.0	0.0	10
Easier	1–2	88.9	9.2	1.8	0.2	7,903
Large	3–4	80.0	17.0	2.7	0.3	2,241
	5 & Up	63.6	27.3	9.1	0.0	11
Easier	1–2	93.0	5.9	1.0	0.1	7,911
Medium	3–4	86.2	11.7	2.0	0.1	2,255
	5 & Up	72.7	18.2	9.1	0.0	11
Easier	1–2	95.0	4.2	0.7	0.1	7,914
Small	3–4	90.3	8.1	1.5	0.1	2,250
	5 & Up	81.8	18.2	0.0	0.0	11
Harder	1–2	93.2	5.9	0.8	0.1	7,970
Lg, Med, Sm	3–4	85.7	12.4	1.7	0.3	2,290
	5 & Up	66.7	33.3	0.0	0.0	12
Harder	1–2	88.8	9.5	1.5	0.2	7,970
Large	3–4	79.6	17.2	2.8	0.4	2,283
	5 & Up	61.5	38.5	0.0	0.0	13
Harder	1–2	93.5	5.6	0.8	0.1	7,965
Medium	3–4	87.4	11.0	1.4	0.3	2,263
	5 & Up	75.0	25.0	0.0	0.0	12
Harder	1–2	95.5	3.9	0.6	0.1	7,967
Small	3–4	91.3	7.5	1.1	0.2	2,264
	5 & Up	100.0	0.0	0.0	0.0	12

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

TABLE 15

Distribution of absolute impact¹ by number of items with shifted-b parameters—Items pretested in positions 1–10 but appeared in positions 21–28 of operational CATs (AC)—Simulated test takers who received studied item(s) in AC positions for both baseline and shifted-b conditions

Direction and Degree of b Shift	Number of Items With b Shift	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Easier Lg, Med, Sm	1–2	93.8	6.0	0.2	0.0	30,536
	3–4	87.3	12.0	0.6	0.0	9,338
	5 & Up	85.0	14.5	0.5	0.0	433
Easier Large	1–2	90.9	8.7	0.4	0.0	30,597
	3–4	81.9	16.8	1.2	0.1	9,277
	5 & Up	74.1	24.0	1.6	0.2	433
Easier Medium	1–2	94.0	5.8	0.2	0.0	30,550
	3–4	87.7	11.6	0.7	0.0	9,324
	5 & Up	84.5	15.0	0.5	0.0	433
Easier Small	1–2	95.5	4.3	0.1	0.0	30,531
	3–4	91.2	8.3	0.5	0.0	9,344
	5 & Up	90.5	9.3	0.2	0.0	432
Harder Lg, Med, Sm	1–2	94.2	5.6	0.2	0.0	30,471
	3–4	88.6	10.7	0.6	0.1	9,390
	5 & Up	85.0	13.7	1.3	0.0	446
Harder Large	1–2	91.9	7.8	0.3	0.0	30,438
	3–4	83.4	15.2	1.3	0.1	9,418
	5 & Up	76.5	21.3	2.2	0.0	451
Harder Medium	1–2	94.8	5.0	0.2	0.0	30,455
	3–4	89.5	9.7	0.7	0.1	9,405
	5 & Up	86.1	13.0	0.9	0.0	447
Harder Small	1–2	96.1	3.8	0.1	0.0	30,464
	3–4	92.2	7.4	0.4	0.0	9,394
	5 & Up	90.2	9.4	0.4	0.0	449

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

Once again these results show that shifted-b items that were delivered earlier in CATs resulted in more impact. Simulated test takers who encountered the shifted-b items in the last third of their CATs (in the AC positions) were the least affected, while those who encountered the shifted-b items in the first third of their CATs (in the BA positions) experienced the most impact. The general trend from these three tables is quite clear. While the majority of the simulated test takers who encountered shifted-b items experienced no impact or a small score impact, a small group of simulated test takers experienced a score impact of 4 points or more when they received shifted-b items earlier in their CATs. Furthermore, the percent of scores that were affected increased as the number of shifted-b items went up.

What is remarkable is that between 50% and 96% of the simulated test takers who received shifted-b items did not experience impact at all across all of the ability levels and conditions. It is hard to provide an explanation as to why this occurred. This may involve rather complex interactions among various factors, such as the ability level of the simulated test takers, the specific shifted-b items the simulated test takers saw, and when and where these simulated test takers encountered the shifted-b items as well as the other items that were given during their testing sessions.

Routing Effects and Impact

One of the research questions for this project concerned how much a given CAT was altered by the introduction of multiple shifted-b items. How did the algorithm react to this kind of intrusion? To answer this question, the baseline and the shifted-b CATs were compared. Any difference in the test after a shifted-b item was introduced was classified as a routing effect. Two kinds of routing effects were examined in this section. A *fresh item* routing effect occurred when one or more of the items that were administered during the shifted-b CAT sessions were not delivered in the corresponding baseline CATs. A *different order* routing effect occurred when simulated test takers were delivered the same collections of items as in their baseline CATs but in different order.

Analyses were performed to examine the joint distributions of absolute impact by the number of fresh items that simulated test takers received for all the data conditions. Tables 16 through 21 provide such distributions under the easy shift as well as the hard shift data conditions for the three delivery position groups. The first columns of these tables show the direction and the degree of the shift. The second columns display the number of fresh items that simulated test takers encountered in their CAT sessions. The number of fresh items is grouped into seven categories ranging from zero through the maximum remaining test length. A fresh item routing effect of zero indicates that a shifted-b CAT contains the same items as its corresponding baseline CAT.

TABLE 16

Distribution of absolute impact¹ by number of fresh items received²—Items pretested in positions 11–20 but appeared in positions 1–10 of operational CATs (BA)—Simulated test takers who received studied item(s) in BA positions for both baseline and shifted-b conditions, items became easier

Direction and Degree of b Shift	Number of Fresh Items Received	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Easier	0	96.3	3.5	0.2	0.0	62,742
1/3 Large,	1–5	28.2	66.0	5.5	0.2	2,131
1/3 Medium,	6–10	20.3	65.0	13.7	1.1	2,512
& 1/3 Small	11–15	15.4	55.7	24.6	4.3	1,950
Shift	16–20	10.3	36.9	33.3	19.5	604
	21–25	3.9	19.6	41.2	35.3	51
	26–28	0.0	0.0	0.0	100.0	1
Easier	0	95.5	4.2	0.3	0.0	59,286
Large Shift	1–5	28.2	66.3	5.3	0.3	3,174
	6–10	20.4	64.8	13.5	1.2	4,090
	11–15	14.6	55.8	24.9	4.7	3,525
	16–20	7.8	37.7	35.8	18.6	1,134
	21–25	2.7	23.0	37.8	36.5	74
	26–28	0.0	0.0	0.0	100.0	1
Easier	0	96.2	3.6	0.2	0.0	62,409
Medium	1–5	28.0	66.4	5.3	0.3	2,211
Shift	6–10	20.1	66.0	12.9	1.0	2,671
	11–15	15.0	56.1	24.5	4.4	2,243
	16–20	9.9	37.4	35.6	17.1	689
	21–25	2.3	25.6	34.9	37.2	43
	26–28	0.0	0.0	0.0	100.0	1
Easier	0	96.5	3.3	0.2	0.0	63,988
Small Shift	1–5	26.7	67.2	5.8	0.3	1,720
	6–10	20.3	65.6	13.1	0.9	1,938
	11–15	15.2	56.9	23.8	4.1	1,586
	16–20	10.9	36.5	33.2	19.4	458
	21–25	3.2	25.8	35.5	35.5	31
	26–28	0.0	0.0	0.0	100.0	1

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

²Items that were not delivered in the baseline CAT.

TABLE 17

Distribution of absolute impact¹ by number of fresh items received²—Items pretested in positions 11–20 but appeared in positions 1–10 of operational CATs (BA)—Simulated test takers who received studied item(s) in BA positions for both baseline and shifted-b conditions, items became harder

Direction and Degree of b Shift	Number of Fresh Items Received	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Harder	0	96.2	3.7	0.1	0.0	62,269
1/3 Large, 1/3 Medium, & 1/3 Small Shift	1–5	29.3	66.0	4.6	0.1	1,852
	6–10	21.2	66.5	11.5	0.7	1,799
	11–15	15.0	54.5	24.4	6.0	1,097
	16–20	9.1	42.0	34.9	13.9	352
	21–25	0.0	20.0	53.3	26.7	30
	26–28	0.0	50.0	0.0	50.0	2
Harder Large Shift	0	95.5	4.3	0.1	0.0	57,650
Large Shift	1–5	29.5	66.5	4.0	0.0	2,772
	6–10	21.4	65.8	12.3	0.5	2,945
	11–15	15.1	55.6	24.2	5.1	1,998
	16–20	8.9	39.8	35.5	15.8	620
	21–25	0.0	20.5	56.4	23.1	39
	26–28	0.0	50.0	0.0	50.0	2
Harder Medium Shift	0	96.2	3.7	0.1	0.0	61,713
Medium Shift	1–5	27.6	67.8	4.5	0.1	1,867
	6–10	21.8	65.7	11.7	0.8	1,926
	11–15	14.3	55.5	25.1	5.1	1,225
	16–20	8.8	39.7	36.3	15.2	375
	21–25	0.0	26.1	47.8	26.1	23
	26–28	0.0	0.0	0.0	100.0	1
Harder Small Shift	0	96.5	3.4	0.1	0.0	63,775
Small Shift	1–5	26.2	68.3	5.4	0.1	1,440
	6–10	21.6	65.6	11.9	0.9	1,400
	11–15	14.8	55.2	24.3	5.7	863
	16–20	8.1	42.3	36.9	12.7	260
	21–25	0.0	23.1	53.8	23.1	13
	26–28	0.0	0.0	0.0	100.0	1

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

²Items that were not delivered in the baseline CAT.

TABLE 18

Distribution of absolute impact¹ by number of fresh items received²—Items pretested in positions 1–10 but appeared in positions 11–20 of operational CATs (AB)—Simulated test takers who received studied item(s) in AB positions for both baseline and shifted-b conditions, items became easier

Direction and Degree of b Shift	Number of Fresh Items Received	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Easier	0	98.6	1.3	0.1	0.0	9,236
1/3 Large,	1–5	19.1	72.8	8.0	0.1	703
1/3 Medium,	6–10	10.1	55.6	31.7	2.6	189
& 1/3 Small	11–15	11.8	52.9	23.5	11.8	17
Shift	16–20	0.0	0.0	100.0	0.0	1
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Easier	0	97.7	2.2	0.1	0.0	8,802
Large Shift	1–5	18.8	71.4	9.3	0.5	1,043
	6–10	9.4	55.7	30.3	4.5	287
	11–15	9.1	45.5	27.3	18.2	22
	16–20	0.0	0.0	100.0	0.0	1
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Easier	0	98.5	1.4	0.1	0.0	9,302
Medium	1–5	18.9	73.1	7.9	0.1	683
Shift	6–10	8.5	54.5	33.5	3.4	176
	11–15	13.3	46.7	26.7	13.3	15
	16–20	0.0	0.0	100.0	0.0	1
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Easier	0	99.0	1.0	0.1	0.0	9,545
Small Shift	1–5	20.7	70.4	8.7	0.2	497
	6–10	10.0	58.3	29.2	2.5	120
	11–15	16.7	41.7	25.0	16.7	12
	16–20	0.0	0.0	100.0	0.0	1
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

²Items that were not delivered in the baseline CAT.

TABLE 19

Distribution of absolute impact¹ by number of fresh items received²—Items pretested in positions 1–10 but appeared in positions 11–20 of operational CATs (AB)—Simulated test takers who received studied item(s) in AB positions for both baseline and shifted-b conditions, items became harder

Direction and Degree of b Shift	Number of Fresh Items Received	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Harder	0	98.4	1.5	0.1	0.0	9,398
1/3 Large, 1/3 Medium, & 1/3 Small Shift	1–5	17.8	74.1	7.6	0.5	657
	6–10	13.4	62.4	19.8	4.5	202
	11–15	13.3	33.3	40.0	13.3	15
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Harder	0	97.6	2.3	0.1	0.0	8,891
Large Shift	1–5	17.8	72.5	9.4	0.3	1,065
	6–10	11.3	59.6	23.8	5.3	282
	11–15	7.4	44.4	40.7	7.4	27
	16–20	0.0	0.0	0.0	100.0	1
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Harder	0	98.5	1.5	0.1	0.0	9,449
Medium Shift	1–5	18.0	73.8	7.7	0.5	611
	6–10	12.2	61.0	21.3	5.5	164
	11–15	12.5	37.5	37.5	12.5	16
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Harder	0	98.8	1.1	0.1	0.0	9,710
Small Shift	1–5	17.8	74.1	7.6	0.5	410
	6–10	11.6	58.9	24.1	5.4	112
	11–15	9.1	27.3	45.5	18.2	11
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

²Items that were not delivered in the baseline CAT.

TABLE 20

Distribution of absolute impact¹ by number of fresh items received²—Items pretested in positions 1–10 but appeared in positions 21–28 of operational CATs (AC)—Simulated test takers who received studied item(s) in AC positions for both baseline and shifted-b conditions, items became easier

Direction and Degree of b Shift	Number of Fresh Items Received	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Easier	0	94.1	5.8	0.1	0.0	39,387
1/3 Large,	1–5	12.2	77.6	9.6	0.7	919
1/3 Medium,	6–10	0.0	100.0	0.0	0.0	1
& 1/3 Small	11–15	0.0	0.0	0.0	0.0	0
Shift	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Easier	0	92.0	7.9	0.1	0.0	38,629
Large Shift	1–5	12.0	76.0	11.3	0.7	1,674
	6–10	0.0	100.0	0.0	0.0	4
	11–15	0.0	0.0	0.0	0.0	0
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Easier	0	94.5	5.4	0.1	0.0	39,297
Medium	1–5	11.3	77.8	10.5	0.4	1,009
Shift	6–10	0.0	100.0	0.0	0.0	1
	11–15	0.0	0.0	0.0	0.0	0
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Easier	0	95.9	4.1	0.1	0.0	39,635
Small Shift	1–5	12.1	76.8	10.6	0.6	671
	6–10	0.0	100.0	0.0	0.0	1
	11–15	0.0	0.0	0.0	0.0	0
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

²Items that were not delivered in the baseline CAT.

TABLE 21

Distribution of absolute impact¹ by number of fresh items received²—Items pretested in positions 1–10 but appeared in positions 21–28 of operational CATs (AC)—Simulated test takers who received studied item(s) in AC positions for both baseline and shifted-b conditions, items became harder

Direction and Degree of b Shift	Number of Fresh Items Received	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Harder	0	94.6	5.3	0.1	0.0	39,406
1/3 Large, 1/3 Medium, & 1/3 Small Shift	1–5	13.7	75.9	9.6	0.9	900
	6–10	0.0	0.0	100.0	0.0	1
	11–15	0.0	0.0	0.0	0.0	0
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Harder Large Shift	0	92.9	7.0	0.1	0.0	38,709
Large Shift	1–5	12.7	74.9	11.7	0.7	1,593
	6–10	0.0	60.0	40.0	0.0	5
	11–15	0.0	0.0	0.0	0.0	0
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Harder Medium Shift	0	95.4	4.5	0.1	0.0	39,349
Medium Shift	1–5	13.7	75.7	9.8	0.8	957
	6–10	0.0	0.0	100.0	0.0	1
	11–15	0.0	0.0	0.0	0.0	0
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Harder Small Shift	0	96.4	3.5	0.1	0.0	39,675
Small Shift	1–5	13.6	76.7	8.7	0.9	632
	6–10	0.0	0.0	0.0	0.0	0
	11–15	0.0	0.0	0.0	0.0	0
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

²Items that were not delivered in the baseline CAT.

The results in these tables clearly indicate that simulated test takers experienced more impact as the number of fresh items they received increased and when the shifted-b items were administered earlier in their CAT sessions. They also show that simulated test takers encountered more fresh items when shifted-b items were introduced earlier in their test sessions. When studied items were made easier in the BA positions with a large shift, 10% of the simulated test takers saw up to 10 fresh items, and an additional 7% of the simulated test takers encountered 11 to 28 fresh items. Yet for the items made easier in the AC positions, only 2% of the simulated test takers saw up to 10 fresh items. When items were manipulated to be harder in the BA positions with a large shift, 8% of the simulated test takers were delivered up to 10 fresh items, and another 4% of the simulated test takers saw 11 to 28 fresh items. Only 2% of the simulated test takers in the AC positions were administered up to 10 fresh items.

All eight data conditions of the BA delivery positions produced more impact than the AB and AC positions. For example, in Table 16 under the large shifted-b condition, the percentages of simulated test takers who experienced no impact decreased from 96% when no fresh items were delivered to 3% when 21 to 25 fresh items were administered. The percentages of simulated test takers who saw score changes in the range of five or more score points (large impact) increased from 0.0% when no fresh items were delivered to 37% when 21 to 25 fresh items were encountered. What was not expected is that even the small shift conditions produced similar score impact. The results in these tables also show that hard shift data conditions tend to produce less impact than easy shift conditions.

Tables 22 through Table 27 provide the joint distributions of absolute impact by the number of items that were administered in different order for all the data conditions. A different order routing effect of zero means that a shifted-b CAT contained the same collection of items as its corresponding baseline CAT. In addition, these items were delivered in the same order as its baseline CAT.

TABLE 22

Distribution of absolute impact¹ by number of items received in an order different than in baseline CAT—Items pretested in positions 11–20 but appeared in positions 1–10 of operational CATs (BA) simulated test takers who received studied item(s) in BA positions for both baseline and shifted-b conditions, items became easier

Direction and Degree of b Shift	Number of Items in Different Order	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Easier	0	96.5	3.3	0.2	0.0	62,543
1/3 Large,	1–5	31.0	65.6	3.2	0.2	1,092
1/3 Medium,	6–10	26.0	65.1	8.5	0.4	1,640
& 1/3 Small	11–15	18.5	60.9	18.7	2.0	1,827
Shift	16–20	15.6	56.9	21.6	5.8	2,041
	21–25	16.0	48.8	25.0	10.2	836
	26–28	0.0	33.3	16.7	50.0	12
Easier	0	95.8	4.0	0.2	0.0	59,004
Large Shift	1–5	30.9	65.6	3.2	0.2	1,580
	6–10	25.4	65.6	8.6	0.4	2,453
	11–15	18.9	61.4	17.4	2.3	2,880
	16–20	15.5	56.0	22.8	5.7	3,756
	21–25	13.3	50.3	26.0	10.4	1,599
	26–28	0.0	33.3	16.7	50.0	12
Easier	0	96.4	3.5	0.2	0.0	62,209
Medium	1–5	29.7	66.7	3.4	0.2	1,108
Shift	6–10	25.8	65.9	7.9	0.4	1,641
	11–15	19.0	61.5	17.4	2.1	1,877
	16–20	15.2	57.5	22.0	5.3	2,436
	21–25	16.0	50.2	24.8	8.9	986
	26–28	0.0	40.0	20.0	40.0	10
Easier	0	96.6	3.2	0.2	0.0	63,826
Small Shift	1–5	28.0	68.0	3.8	0.2	851
	6–10	26.0	65.6	8.1	0.3	1,249
	11–15	18.3	62.1	17.6	2.0	1,369
	16–20	16.4	57.4	21.0	5.2	1,718
	21–25	16.0	51.6	23.3	9.1	701
	26–28	0.0	37.5	25.0	37.5	8

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

TABLE 23

Distribution of absolute impact¹ by number of items received in an order different than in baseline CAT—Items pretested in positions 11–20 but appeared in positions 1–10 of operational CATs (BA)—Simulated test takers who received studied item(s) in BA positions for both baseline and shifted-b conditions, items became harder

Direction and Degree of b Shift	Number of Items in Different Order	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Harder	0	96.4	3.5	0.1	0.0	62,090
1/3 Large, 1/3 Medium, & 1/3 Small Shift	1–5	33.2	64.1	2.6	0.1	961
	6–10	24.2	67.7	7.8	0.2	1,376
	11–15	20.4	63.3	14.5	1.8	1,350
	16–20	17.8	55.8	20.2	6.2	1,187
	21–25	12.0	49.8	30.3	8.0	426
	26–28	0.0	27.3	36.4	36.4	11
Harder	0	95.8	4.1	0.1	0.0	57,401
Large Shift	1–5	33.0	65.2	1.7	0.0	1,374
	6–10	25.3	66.9	7.7	0.1	2,109
	11–15	19.8	63.6	15.2	1.4	2,120
	16–20	18.1	55.7	20.7	5.5	2,228
	21–25	11.5	51.1	28.7	8.7	783
	26–28	0.0	27.3	36.4	36.4	11
Harder	0	96.4	3.5	0.1	0.0	61,527
Medium Shift	1–5	31.2	66.2	2.6	0.1	937
	6–10	23.9	67.9	7.9	0.2	1,388
	11–15	20.5	63.4	14.7	1.4	1,387
	16–20	17.5	56.4	20.6	5.5	1,412
	21–25	12.1	50.3	28.9	8.7	471
	26–28	0.0	25.0	37.5	37.5	8
Harder	0	96.6	3.2	0.1	0.0	63,640
Small Shift	1–5	29.6	67.1	3.2	0.1	717
	6–10	22.8	68.3	8.6	0.2	1,033
	11–15	20.2	63.4	14.9	1.4	1,038
	16–20	17.7	57.4	19.3	5.6	995
	21–25	12.1	49.7	30.4	7.8	322
	26–28	0.0	14.3	57.1	28.6	7

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

TABLE 24

Distribution of absolute impact¹ by number of items received in an order different than in baseline CAT—Items pretested in positions 1–10 but appeared in positions 11–20 of operational CATs (AB) simulated test takers who received studied item(s) in AB positions for both baseline and shifted-b conditions, items became easier

Direction and Degree of b Shift	Number of Items in Different Order	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Easier	0	98.7	1.2	0.1	0.0	9,214
1/3 Large,	1–5	20.3	72.7	7.0	0.0	597
1/3 Medium,	6–10	13.3	62.9	22.3	1.5	264
& 1/3 Small	11–15	10.8	58.5	26.2	4.6	65
Shift	16–20	0.0	33.3	50.0	16.7	6
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Easier	0	98.0	1.9	0.1	0.0	8,765
Large Shift	1–5	19.1	72.5	8.1	0.3	908
	6–10	13.7	59.8	23.3	3.1	386
	11–15	10.2	55.7	28.4	5.7	88
	16–20	0.0	37.5	37.5	25.0	8
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Easier	0	98.6	1.3	0.1	0.0	9,282
Medium	1–5	20.4	72.7	6.9	0.0	583
Shift	6–10	11.6	62.0	24.4	2.0	250
	11–15	10.9	58.2	25.5	5.5	55
	16–20	0.0	42.9	42.9	14.3	7
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Easier	0	99.0	0.9	0.1	0.0	9,531
Small Shift	1–5	22.4	69.7	7.9	0.0	419
	6–10	12.8	66.1	20.0	1.1	180
	11–15	15.0	52.5	25.0	7.5	40
	16–20	0.0	20.0	60.0	20.0	5
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

TABLE 25

Distribution of absolute impact¹ by number of items received in an order different than in baseline CAT—Items pretested in positions 1–10 but appeared in positions 11–20 of operational CATs (AB)—Simulated test takers who received studied item(s) in AB positions for both baseline and shifted-b conditions, items became harder

Direction and Degree of b Shift	Number of Items in Different Order	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Harder	0	98.6	1.3	0.1	0.0	9,374
1/3 Large, 1/3 Medium, & 1/3 Small Shift	1–5	18.3	75.3	6.2	0.2	546
	6–10	13.8	65.5	17.1	3.6	275
	11–15	17.6	58.1	20.3	4.1	74
	16–20	66.7	33.3	0.0	0.0	3
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Harder	0	97.9	2.1	0.1	0.0	8,855
Large Shift	1–5	19.0	73.3	7.6	0.1	902
	6–10	12.4	64.1	19.4	4.0	396
	11–15	13.3	55.2	28.6	2.9	105
	16–20	25.0	37.5	25.0	12.5	8
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Harder	0	98.6	1.3	0.1	0.0	9,425
Medium Shift	1–5	18.6	74.6	6.6	0.2	515
	6–10	13.8	65.7	16.3	4.2	239
	11–15	15.5	53.4	25.9	5.2	58
	16–20	33.3	66.7	0.0	0.0	3
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Harder	0	98.9	1.0	0.1	0.0	9,694
Small Shift	1–5	19.3	74.5	6.2	0.0	353
	6–10	11.8	65.4	18.3	4.6	153
	11–15	14.0	48.8	30.2	7.0	43
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

TABLE 26

Distribution of absolute impact¹ by number of items received in an order different than in baseline CAT—Items pretested in positions 1–10 but appeared in positions 21–28 of operational CATs (AC)—Simulated test takers who received studied item(s) in AC positions for both baseline and shifted-b conditions, items became easier

Direction and Degree of b Shift	Number of Items in Different Order	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Easier	0	94.2	5.7	0.1	0.0	39,341
1/3 Large,	1–5	12.7	77.4	9.3	0.6	961
1/3 Medium,	6–10	0.0	80.0	20.0	0.0	5
& 1/3 Small	11–15	0.0	0.0	0.0	0.0	0
Shift	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Easier	0	92.1	7.8	0.1	0.0	38,546
Large Shift	1–5	12.5	75.8	11.0	0.7	1,750
	6–10	0.0	90.9	9.1	0.0	11
	11–15	0.0	0.0	0.0	0.0	0
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Easier	0	94.6	5.3	0.1	0.0	39,242
Medium	1–5	11.9	77.4	10.3	0.4	1,059
Shift	6–10	0.0	83.3	16.7	0.0	6
	11–15	0.0	0.0	0.0	0.0	0
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Easier	0	95.9	4.0	0.1	0.0	39,601
Small Shift	1–5	12.7	76.5	10.3	0.6	701
	6–10	0.0	80.0	20.0	0.0	5
	11–15	0.0	0.0	0.0	0.0	0
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

TABLE 27

Distribution of absolute impact¹ by number of items received in an order different than in baseline CAT—Items pretested in positions 1–10 but appeared in positions 21–28 of operational CATs (AC)—Simulated test takers who received studied item(s) in AC positions for both baseline and shifted-b conditions, items became harder

Direction and Degree of b Shift	Number of Items in Different Order	Absolute Difference in Estimated True Scores				N
		0 %	1–2 %	3–4 %	5 & Up %	
Harder	0	94.7	5.3	0.1	0.0	39365
1/3 Large, 1/3 Medium, & 1/3 Small Shift	1–5	14.0	75.9	9.3	0.9	939
	6–10	0.0	33.3	66.7	0.0	3
	11–15	0.0	0.0	0.0	0.0	0
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Harder	0	93.0	6.8	0.1	0.0	38626
Large Shift	1–5	13.2	74.8	11.3	0.7	1672
	6–10	11.1	55.6	33.3	0.0	9
	11–15	0.0	0.0	0.0	0.0	0
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Harder	0	95.5	4.5	0.1	0.0	39307
Medium Shift	1–5	14.1	75.6	9.5	0.8	996
	6–10	25.0	25.0	50.0	0.0	4
	11–15	0.0	0.0	0.0	0.0	0
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0
Harder	0	96.5	3.5	0.0	0.0	39649
Small Shift	1–5	13.9	76.7	8.5	0.9	656
	6–10	0.0	50.0	50.0	0.0	2
	11–15	0.0	0.0	0.0	0.0	0
	16–20	0.0	0.0	0.0	0.0	0
	21–25	0.0	0.0	0.0	0.0	0
	26–28	0.0	0.0	0.0	0.0	0

¹Absolute difference between Estimated True Score Baseline Condition and Estimated True Score Shifted-b Condition.

The results of different order routing effects in these tables show similar patterns as those described for fresh items routing effects. More shifted-b CATs were delivered in an order that was different than that of the baseline CATs when studied items were introduced earlier in the test. As expected, as simulated test takers received more items delivered in an order that was different from the baseline CATs, they experienced more impact. This pattern holds across delivery positions, shift directions, as well as degree of shift. The easy shift condition seemed to produce more impact than the hard shift condition. Simulated test takers experienced more impact in the BA delivery positions than in the AB and AC positions. Different order routing effect has less impact on scores than fresh item routing effect. In fact, at least twice as many simulated test takers received scores different from their baseline scores by five points or more when they received 11 to 25 fresh items rather than 11 to 25 items that were delivered in a different order. Even so, the impact of different order routing effect on ability estimation cannot be disregarded because 1% to 10% of the simulated test takers experienced a score change of five points or more when they received 11 to 25 items that were administered in a different order.

The results of routing effects and impact are not unexpected. The flawed-item study by Liu and Steffen (1999) found that the CAT algorithm did try to self-adapt after encountering a flawed item early in the CAT session.

Discussion

The results indicate that studied items with a large shift produced more score impact than the other three levels of parameter shift. In addition, more impact was observed when the studied items were made easier than when these items were made harder. Furthermore, simulated test takers were likely to experience more impact as the number of shifted-b items they encountered increased. This is particularly evident when the shifted-b items appeared early in a test. Finally, early appearance of the shifted-b items led to large routing effects, which in turn seemed to produce more scores that differed from their baseline scores by five points or more.

Even though these findings demonstrate that multiple items with shifted-b parameters did not result in impact as severe as that produced by a single flawed item (see Liu & Steffen, 1999), accumulation of smaller

effects from these items could still adversely impact some test taker scores. Moreover, the impact was not uniform across the data conditions investigated in this study. The severity of the impact was influenced by the complex interactions of the ability level of the test takers, the specific shifted-b items they saw, the number of shifted-b items they received, and when and where they encountered these items as well as the other items administered during their testing sessions. Such differential effect, albeit affecting a very small group of test takers, could be problematic if it occurred operationally.

One of the solutions for operational CAT programs is to implement quality control systems to routinely monitor the test data and try to detect test takers' item responding behavior that is affected by context effects (such as item location effects). Whether such behavior can be detected using observed response data remains to be seen. Remember, response data from a CAT program are sparse since not every test taker receives the same collection of items. This presents serious problems for data analysis and evaluation. Successful detection of abnormal item response behavior is only half the battle. Now a testing program has to decide its course of action concerning test takers that are affected by such abnormal behavior. Currently there are no known industry standards and procedures that can be used to cope with this issue. There is no doubt that such a void needs to be filled in order for CAT to survive as a viable venue for high-stakes testing programs. Another solution is to develop a CAT algorithm that incorporates parameters such as item delivery positions.

While considering the results of this study, readers need to keep in mind that the results should not be generalized to settings that are different from those investigated in the study. For example, the results reported here dealt only with impact from one direction at a time. That is, the studied items were made either easier or harder. If a test taker received several such items in a test, the impact on his/her score would be cumulative. In a scenario where some items become easier and some harder due to item location effects, a test taker might encounter both types of items in one test session. In this case, some of the item location effects might be cancelled out to produce a smaller impact than those reported here. In a sense, this study describes the worst case scenarios that may be observed if the selected magnitudes of parameter shifts occurred in an operational setting. These magnitudes of parameter shifts, however, are quite conservative. Such levels are similar to those observed in a paper-and-pencil testing environment in which item position is held relatively constant. In an actual CAT environment in which item position is typically not constrained, these levels of parameter shift probably represent the lower bounds.

Another caveat to keep in mind is that this is a simulation study. No matter how we tried to mimic the real world settings, we could not simulate accurately how test takers would interact with particular aspects of the testing environment in which they were administered items whose parameter estimates did not hold due to change in delivery positions. Our simulation study simply provides a very general direction, and we hope others will join us to further examine the impact of item location effects and other context effects on CAT scores using real data from a controlled experimental environment as well as operational settings.

Note that the research presented here relates directly to computerized item-by-item adaptive testing environments. When using more constrained adaptive testing procedures, context effects such as item position can be mitigated or even kept constant. For example, within the multiple-form structure (MFS) methodology that is being investigated at LSAC, intact forms can be calibrated (or tried out) before they are delivered operationally, much as the current paper-and-pencil forms are.

References

- Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, 5(3), 225–264.
- Eignor, D. R. (1985). *An investigation of the feasibility and practical outcomes of preequating the SAT verbal and mathematical sections* (Research Report No. 85-10). Princeton, NJ: Educational Testing Service.
- Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, 8, 147–154.
- Kolen, M. J., & Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement*, 27(1), 27–39.
- Liu, M., & Steffen, M. (1999). *The impact of flawed items on ability estimation in CAT*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.
- Stocking, M. L., & Lewis C. (1995). *Controlling item exposure conditional on ability in computerized adaptive testing* (Research Report No. 93-2). Princeton, NJ: Educational Testing Service.

-
- Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277–292.
- Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*, 151–166.
- Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10*(3), 10–15.

Authors Note

Mei Liu is now at ETS.

Renbang Zhu is now at the American Board of Internal Medicine.

Fanmin Guo is now at The Graduate Management Admission Council.